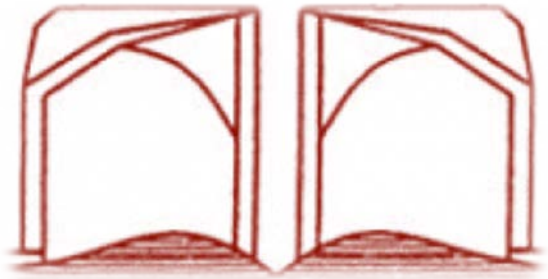


روش‌های یادگیری ماشین در پردازش زبان طبیعی

مروری بر آمار و تخمین

هادی ویسی

h.veisi@ut.ac.ir

دانشگاه تهران – دانشکده علوم و فنون نوین

فهرست

احتمال . . .

- بیان میزان اطمینان از خروجی وقایعی (مشاهده‌هایی) که قطعی نیستند

- مثال

- در پرتاب یک سکه، شانس (احتمال) آمدن شیر؟
  - دو حالت داریم (شیر یا خط)
  - اگر سکه سالم باشد، شانس ۵۰٪-۵۰٪ است، پس احتمال آمدن شیر=۱/۲

- در پرتاب یک تاس، شانس (احتمال) آمدن عدد شش؟
  - شش حالت داریم (۱ تا ۶)
  - شانس (احتمال) آمدن عدد شش = ۱/۶

- در یک متن ۲۰۰ کلمه‌ای، «سرزمین» ۳ بار تکرار شده است.
  - شانس (احتمال) آمدن «سرزمین» در این متن = ۳/۲۰۰

مفهوم ایران بزرگ از جهات گوناگون ریشه در تاریخ چند هزار ساله آن دارد و به دوران نخستین امپراتوری ایرانی که توسط پارس‌ها بنیان گذاشته‌شد بازمی گردد. در دوران جدید، ایران بسیاری از سرزمینهای خود را از دست داد از جمله واگذاری بخش‌های غربی در ترکیه و عراق امروز به امپراتوری عثمانی (۱۵۳۳ میلادی)، واگذاری بخش‌های شرقی در افغانستان امروز به بریتانیا طی قرار داد پاریس در ۱۸۵۷ میلادی و ۱۹۰۵ میلادی و واگذاری سرزمینهای قفقاز به روسیه در قرن هجدهم و نوزدهم میلاد؛ عهدنامه ترکمانچای در سال ۱۸۲۸ و پس از نبرد روسیه و ایران، استانهای قفقازی ایران را برای همیشه به روسیه واگذار کرد و مرزهای جدید در طول رودخانه ارس شکل گرفت. بر طبق عهدنامه گلستان در سال ۱۸۱۳، مناطق ارمنستان، جمهوری آذربایجان و شرق گرجستان که بیشتر بخشی از ایران بودند، به روسیه واگذار شدند. در اثر این تجربه تاریخی کشورها و ملت‌های جدیدی تحت نفوذ روسیه و انگلستان شکل گرفتند که اگرچه از طریق زبان یا فرهنگ با ایران پیوستگی داشتند اما جوامع خاص خود را شکل دادند. در سال ۱۹۳۵ در زمان سلطنت رضا شاه، نام ایران رسماً در مجامع بین‌المللی به‌عنوان نام بخش بجا مانده از سرزمین ایران بکار رفت.

احتمال . . .

تعاریف

- **فضای نمونه** (Sample Space): **مجموعه‌ای از تمام خروجی‌های ممکن** S =
  - کل کلمات در یک پیکره متنی
- **رویداد** (Event): **زیرمجموعه‌ای از فضای نمونه** A=
  - رخداد کلمه‌ای مانند «سرزمین»
- **احتمال** (Probability) **یک رویداد: فراوانی نسبی رخداد آن رویداد با فرض تکرار این فرایند به تعداد دفعات زیاد تحت شرایط مشابه** P(A)=

تعداد مشاهده‌هایی که خروجی آن‌ها متعلق به رویداد A است

$$P(A) = \frac{N_A}{N_S}$$

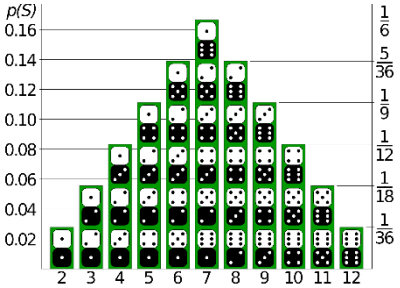
تعداد کل تمام مشاهده‌ها

$$0 \le P(A) \le 1 \quad \text{for all } A$$

$$P_{monogram}(w_i) = \frac{Count(w_i)}{Count(All\,Words)} = \frac{N(w_i)}{N_{total}}$$

احتمال . . .

- **احتمال توأم (Joint Probability)**
  - احتمال دو رویداد A و B که به طور هم‌زمان اتفاق می‌افتند



- مثال: پرتاب هم‌زمان دو تاس، احتمال آمدن جفت شش؟
  - کل حالات = ۶*۶=۳۶ (چون مستقل هستند)
  - احتمال جفت شش = ۱/۳۶

- مثال: رخداد دو کلمه «سرزمین ایران» در یک متن ۲۰۰ کلمه‌ای (Bigram)
  - کل دو کلمه‌ای‌ها؟
    - ۲۰۰
  - احتمال آمدن «سرزمین ایران»؟
    - ۱/۲۰۰؟

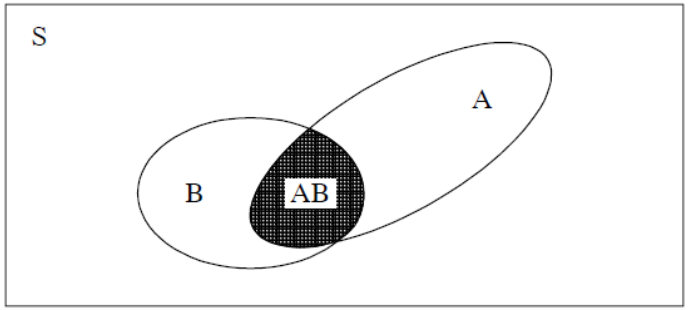
معهوم ایران بزرگ از جهات گوناگون ریشه در تاریخ چند هزار ساله آن دارد و به دوران نخستین امپراتوری ایرانی که توسط پارس‌ها بنیان گذاشته‌شد بازمی گردد. در دوران جدید، ایران بسیاری از سرزمینهای خود را از دست داد از جمله واگذاری بخش‌های غربی در ترکیه و عراق امروز به امپراتوری عثمانی (۱۵۳۳ میلادی)، واگذاری بخش‌های شرقی در افغانستان امروز به بریتانیا طی قرار داد پاریس در ۱۸۵۷ میلادی و ۱۹۰۵ میلادی و واگذاری سرزمینهای قفقاز به روسیه در قرن هجدهم و نوزدهم میلاد؛ عهدنامه ترکمانچای در سال ۱۸۲۸ و پس از نبرد روسیه و ایران، استانهای قفقازی ایران را برای همیشه به روسیه واگذار کرد و مرزهای جدید در طول رودخانه ارس شکل گرفت. بر طبق عهدنامه گلستان در سال ۱۸۱۳، مناطق ارمنستان، جمهوری آذربایجان و شرق گرجستان که بیشتر بخشی از ایران بودند، به روسیه واگذار شدند. در اثر این تجزیه تاریخی کشورها و ملت‌های جدیدی تحت نفوذ روسیه و انگلستان شکل گرفتند که اگرچه از طریق زبان یا فرهنگ با ایران پیوستگی داشتند اما جوامع خاص خود را شکل دادند. در سال ۱۹۳۵ در زمان سلطنت رضا شاه، نام ایران رسماً در مجامع بین‌المللی به‌عنوان نام بخش بجا مانده از **سرزمین ایران** بکار رفت.

$$P(AB) = \frac{N_{AB}}{N_S} \Rightarrow p(w_1, w_2) = \frac{count(w_1, w_2)}{\sum_{w_{1'}, w_{2'}} count(w_{1'}, w_{2'})}$$

احتمال . . .

- **احتمال شرطی (Conditional Probability)**
  - رخ دادن رویداد A با دانستن اینکه رویداد دیگری مانند Bرخ داده است

$$P(A \mid B) = \frac{P(AB)}{P(B)} = \frac{N_{AB}/N_S}{N_B/N_S}$$



- در پردازش متن: احتمال آمدن یک کلمه (w_j) با فرض دانستن کلمه قبلی (w_i)
  - احتمال آمدن «w_j=ایران» بعد از کلمه «w_i=سرزمین»

$$P_{bigram}(w_j \mid w_i) = \frac{N(w_i w_j)}{N(w_i)}$$

احتمال . . .

- **قاعده زنجیری (Chain Rule)**

$$P(A|B) = \frac{P(AB)}{P(B)} = \frac{N_{AB}/N_S}{N_B/N_S} \Rightarrow P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

- **می‌تواند احتمال توأم چندین رویداد را بر حسب ضرب چند احتمال شرطی مشخص کند**
- **استفاده برای تجزیه یک مسئله احتمالاتی توأم پیچیده به زنجیره‌ای از احتمال‌های شرطی**

- **حالت کلی**

$$P(A_1A_2\cdots A_n) = P(A_n|A_1\cdots A_{n-1})\cdots P(A_2|A_1)P(A_1)$$

- **مثال: در پردازش متن**
  - برای دو متغیر: P(A,B) = P(A)P(B|A)
    - P(بازیکنان تیم)=P(بازیکنان|تیم)P
  - برای چهار متغیر: P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)
    - P(بازیکنان تیم ملی نوجوانان)=P(بازیکنان تیم ملی|نوجوانان)P(بازیکنان تیم ملی|تیم)P(بازیکنان تیم|ملی)P

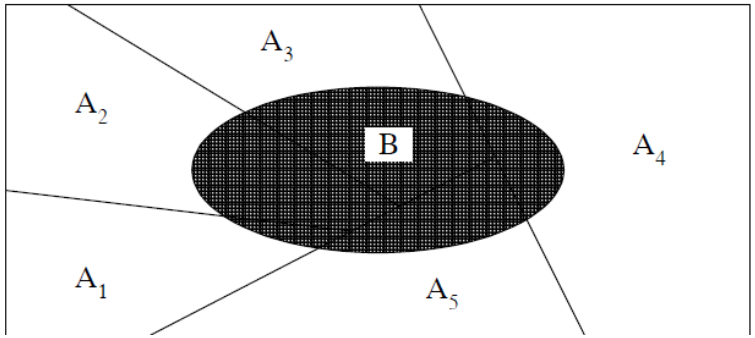
احتمال . . .

- **مستقل (Independent) بودن**
  - **رخ دادن یک رویداد هیچ ارتباط و تأثیری بر رخ دادن رویداد دیگر ندارد.**
    - رخداد A: شیر آمدن یک سکه بعد از پرتاب
    - رخداد B: آمدن شش در پرتاب تاس
  - **احتمال شرطی: برابر با احتمال غیرشرطی است.** $P(A \mid B) = P(A)$
  - **احتمال توأم : برابر با حاصل‌ضرب دو احتمال است.** $P(AB) = P(A)\ P(B)$

احتمال . . .

- افراز (Partition) رویداد B
  - اگر تعداد n رویداد A_1، A_2، ...، A_n یک افراز از S باشد و B یک رویداد در S باشد
  - آنگاه رویدادهای BA_1، BA_2، ...، BA_n یک افراز از B را شکل می‌دهد

$$B = A_1B \cup A_2B \cup \cdots \cup A_nB$$



- چون رویدادهای BA_1، BA_2، ...، BA_n مجزا هستند
  - احتمال رویداد B از حاصل جمع احتمال‌های توأم محاسبه می‌شود

$$P(B) = \sum_{k=1}^{n} P(A_kB)$$

احتمال حاشیه‌ای (Marginal Probability) رویداد B

احتمال

قانون بیز (Bayes´ Rule)

- با توجه به قاعده زنجیری

$$P(B) = \sum_{k=1}^{n} P(A_k)P(B \mid A_k)$$

$$P(A_i \mid B) = \frac{P(A_i B)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum_{k=1}^{n} P(B \mid A_k)P(A_k)}$$

- این قانون مبنای بازشناسی الگوی آماری (مانند بازشناسی گفتار) است

متغیرهای تصادفی . . .

- **متغیر تصادفی (Random Variable)**
  - متغیر X که بیانگر یک کمیت عددی در یک فضای نمونه است
    - عناصر یک فضای نمونه را می‌توان شماره‌گذاری کرد و با آن شماره‌ها به آن‌ها ارجاع کرد.
  - تابعی که هر خروجی ممکن s در فضای نمونه S را به یک عدد حقیقی $X(s)$ نگاشت می‌کند.
  - یک رویداد به‌صورت مجموعه‌ای از $\{s\}$ نشان داده می‌شود که $\{s \mid X(s)=x\}$

- **مثال: پرتاب سکه**
  - فضای نمونه S = {شیر، خط}
  - متغیر تصادفی X

$$X(s) = \begin{cases} 1 & if \quad s = \text{شیر} \\ 0 & if \quad s = \text{خط} \end{cases}$$

- **احتمال اینکه $X=x$ باشد**

$$P(X = x) = P(s \mid X(s) = x)$$

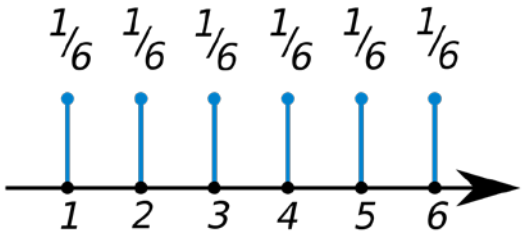
  - در مثال سکه

$$P(X = 1) = P(s|X(s) = 1) = P(s = \text{شیر}) = 0.5$$

متغیرهای تصادفی . . .

- **متغیر تصادفی گسسته (Discrete) . . .**
  - **فقط تعداد متناهی n از مقادیر مختلف را می‌گیرد (دارای توزیع گسسته)**
    - مثال: پرتاب تاس (فقط ۶ حالت دارد)
  - **تابع احتمال (Probability Function)**
    - یا تابع جرم احتمال (Probability Mass Function)
    - برای هر عدد حقیقی x، بیانگر میزان احتمال متغیر تصادفی گسسته است

$$p_X(x) = P(X = x)$$



- مثال: پرتاب تاس
  - متغیرهای تصادفی: 1 تا 6
- حاصل جمع جرم احتمال در تمام مقادیر متغیر تصادفی برابر با یک است

$$\sum_{k=1}^{n} p(x_i) = \sum_{k=1}^{n} P(X = x_i) = 1$$

متغیرهای تصادفی . . .

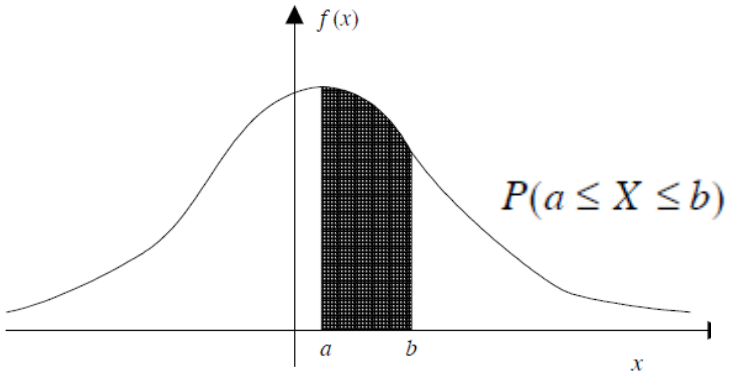
- **متغیر تصادفی پیوسته (Continuous) . . .**
  - **دارای مقادیر پیوسته (و در نتیجه توزیع پیوسته) است**
    - مثال: قد افراد یک کشور، مقدار دامنه سیگنال گفتار
  - **اگر تابع غیرمنفی f وجود داشته باشد که روی مقادیر حقیقی تعریف شده و برای بازه A**

$$P(X \in A) = \int_A f_X(x)dx$$

$$f(x) \geq 0 \text{ for } -\infty \leq x \leq \infty$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

  - **f_x: تابع توزیع احتمال** (pdf: Probability Density Function)
    - احتمال در یک بازه معنی دارد
    - احتمال در یک نقطه برابر با صفر است



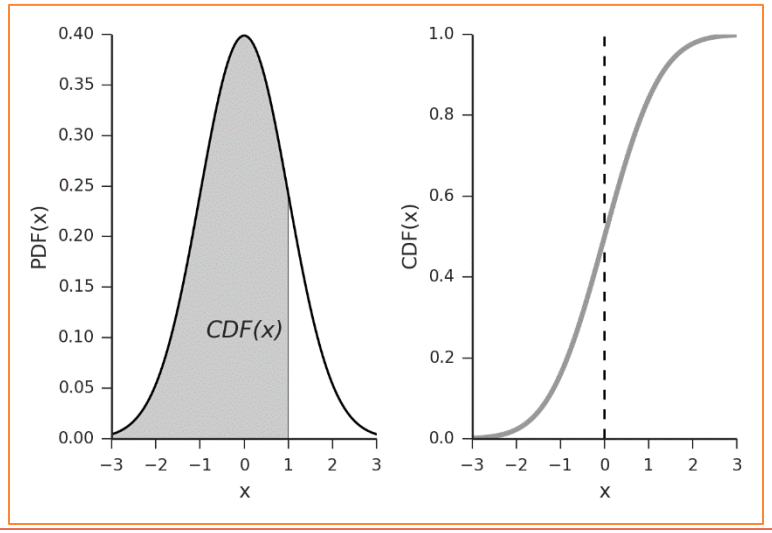
متغیرهای تصادفی . . .

تابع توزیع (Distribution Function)

- یا تابع توزیع تجمعی (CDF: Cumulative Distribution Function)
- بیانگر [جمع] احتمال‌های مقادیر کوچک‌تر از x
- برای متغیر تصادفی گسسته یا پیوسته

$$F(x) = P(X \le x) \quad \text{for} \quad -\infty \le x \le \infty$$

مقدار حقیقی



- برای حالت پیوسته داریم

$$F(x) = \int_{-\infty}^{x} f_X(x)dx$$

$$f_X(x) = \frac{dF(x)}{dx}$$

میانگین و واریانس . . .

- امید ریاضی (Expectation) یا میانگین (Mean) . . .
  - برای متغیر تصادفی گسسته X

$$E(X) = \sum_x xf(x)$$

  - برای متغیر تصادفی پیوسته X

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

  - برای تابعی از متغیر X

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

    - مرکز جرم توزیع احتمال

- امید ریاضی یک عملگر خطی است؛ دارای ویژگی‌های جمع‌پذیری و همگنی
  - حتی در صورت مستقل نبودن X_iها

$$E(a_1X_1 + \cdots + a_nX_n + b) = a_1E(X_1) + \cdots + a_nE(X_n) + b$$

مقدار ثابت

میانگین و واریانس . . .

- امید ریاضی (Expectation) یا میانگین (Mean)
  - مثال ۱: انداختن یک تاس
    - متغیر تصادفی با شش مقدار 1, 2, …,6 با احتمال برابر ۱/۶ برای هر کدام

$$E(X) = \sum_x x f(x)$$

$$= \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \frac{1}{6} \times 3 + \frac{1}{6} \times 4 + \frac{1}{6} \times 5 + \frac{1}{6} \times 6 = 3.5$$

- مثال ۲: محاسبه معدل درسی یک دانشجو
  - متغیر تصادفی (X)؟
    - نمره درس
  - احتمال (تابع توزیع)؟
    - متناسب با تعداد واحد درس‌ها (تعداد واحد درس تقسیم بر کل واحدها)

درس	نمره	تعداد واحد
آواشناسی	۱۸	۲
برنامه نویسی	۱۲	۴
ریاضیات	۱۵	۳
روش تحقیق	۲۰	۱

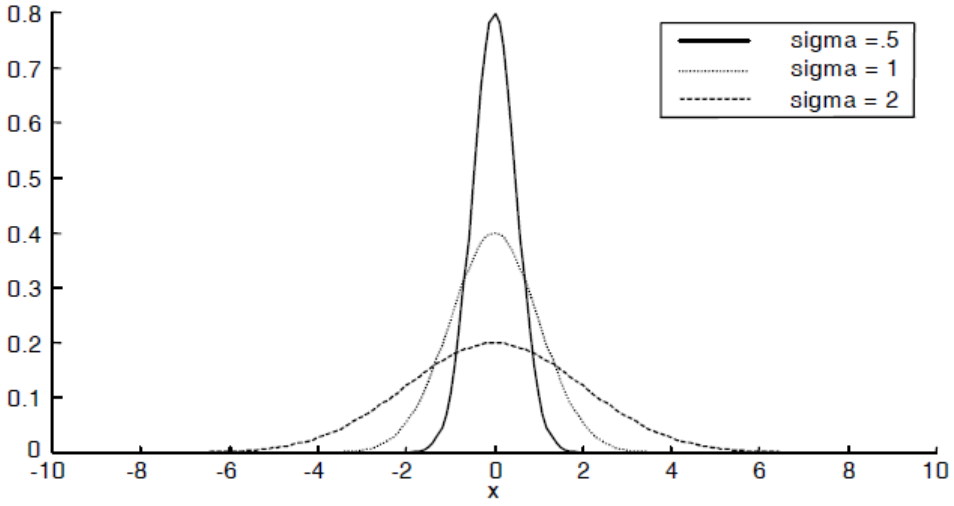
$$E(x) = \frac{2}{10} \times 18 + \frac{4}{10} \times 12 + \frac{3}{10} \times 15 + \frac{1}{10} \times 20 = 14.9$$

میانگین و واریانس . . .

- واریانس (Variance) . . .
  - میانگین متغیر X: $\mu=E(X)$

$$Var(X)=\sigma^2=E\left[(X-\mu)^2\right]$$

- مجذور غیرمنفی واریانس = انحراف معیار (Standard Deviation) σ =
- بیانگر میزان پراکندگی یا انتشار توزیع در اطراف میانگین
  - مقدار کوچک واریانس = توزیع فشرده احتمال در اطراف میانگین
  - مقدار بزرگ واریانس = توزیع احتمال در اطراف میانگین دارای پراکندگی بیشتری است



میانگین و واریانس . . .

دانشجو	نمره
۱	12.9
۲	16.0
۳	12.8
۴	16.1
۵	19.0
۶	16.1
۷	17.4
۸	12.0
۹	12.2
۱۰	15.8

- واریانس (Variance) . . .
  - **مثال: نمرات یک درس**
    - متغیر تصادفی: نمره
    - احتمال؟
      - برابر با ۱/۱۰ (احتمال انتخاب یک دانشجو از بین ۱۰ دانشجو)

$$Var(X) = E((X - E(X))^2) = E(X^2) - E^2(X)$$
$$Var(X) = \sum_x p(x)\ (x - E(X))^2$$

$$E(x) = \mu = \frac{1}{10} \times (12.9 + 16.0 + \cdots + 15.8) = 15$$

$$Var(x) = \sigma^2 = \frac{1}{10} \times \left((12.9 - 15)^2 + (16.0 - 15)^2 + \cdots + (15.8 - 15)^2\right) = 5.8$$

$$\sigma = \sqrt{Var(x)} = \sqrt{5.8} = 2.4$$

میانگین و واریانس . . .

واریانس (Variance)

- ممان (گشتاور) kام X = امید ریاضی $E(X^k)$
  - برای هر متغیر تصادفی Xو هر عدد صحیح مثبت k

- برای واریانس داریم

$$Var(X)=\sigma^2=E\left[(X-\mu)^2\right]=E(X^2)-\left[E(X)\right]^2$$

- واریانس اختلاف بین ممان دوم و مربع ممان اول است

- ویژگی‌های واریانس
  - جمع‌پذیری: در صورت مستقل بودن متغیر تصادفی Xو Y

$$Var(X+Y)=Var(X)+Var(Y)$$

  - همگنی: برقرار نیست

$$Var(aX)=a^2Var(X)$$

  - واریانس مقدار ثابت صفر است

$$Var(a_1X_1+\cdots+a_nX_n+b)=a_1^2Var(X_1)+\cdots+a_n^2Var(X_n)$$

میانگین و واریانس . . .

- امید ریاضی شرطی (Conditional Expectation)
  - برای متغیرهای گسسته X و Y

$$E_{Y|X}(Y \mid X = x) = \sum_{y} y f_{Y|X}(y \mid x)$$

    - امید ریاضی شرطی Y: تابعی از X
  - برای متغیرهای پیوسته X و Y

$$E_{Y|X}(Y \mid X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x) dy$$

- خود $E(Y \mid X)$ یک متغیر تصادفی است
  - تابعی از متغیر تصادفی X است

$$E_X\left[E_{Y|X}(Y \mid X)\right] = E_{X,Y}(Y)$$

- فرض کنید X و Y یک توزیع توأم پیوسته دارند و $g(X,Y)$ تابعی از Xو Y است

$$E_{Y|X}\left[g(X,Y) \mid X = x\right] = \int_{-\infty}^{\infty} g(x,y) f_{Y|X}(y \mid x) dy$$

$$E_X\left\{E_{Y|X}\left[g(X,Y) \mid X\right]\right\} = E_{X,Y}\left[g(X,Y)\right]$$

میانگین و واریانس

میانه (Median)

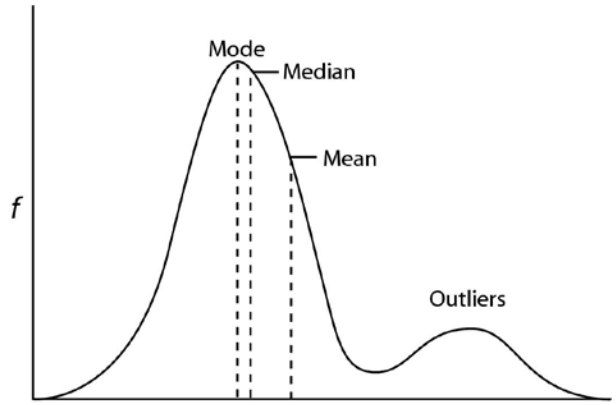
- احتمال کل را به دو قسمت مساوی تقسیم می‌کند
- میانه توزیع متغیر X نقطه‌ای مانند m است

$$P(X \le m) \ge 1/2 \text{ and } P(X \ge m) \ge 1/2$$

  - احتمال سمت چپ m و احتمال سمت راست m دقیقاً 0.5 است

نما (Mode)

- جایی که تابع توزیع دارای بیشترین مقدار خود است
- یک توزیع می‌تواند بیش از یک میانه داشته باشد



قانون اعداد بزرگ . . .

- **قانون اعداد بزرگ (Law of Large Numbers)**
  - **میانگین نمونه (Sample Mean) و واریانس نمونه (Sample Variance)**
    - مقدار میانگین و واریانس تعدادی از نمونه‌های حاصل از یک آزمایش آماری است (آنچه ما در عمل محاسبه می‌کنیم)
- **فرض کنید یک توزیع با میانگین μ و واریانس σ^2 داریم**
  - متغیرهای تصادفی $x_1, x_2, \ldots, x_n$ از این توزیع تولید می‌شوند
  - **متغیرهای تصادفی** i.i.d: - Independent Identically Distributed
    - مستقل و با توزیع یکسان
    - هر یک دارای μ و واریانس σ^2
  - **میانگین حسابی n نمونه**
    - همان میانگین نمونه

$$\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$$

خودش متغیری تصادفی است

قانون اعداد بزرگ

- داریم
  - میانگین "میانگین نمونه"
    - میانگین "میانگین نمونه" برابر با میانگین توزیع است

$$E(\bar{X}_n) = \mu$$

  - واریانس "میانگین نمونه"
    - واریانس "میانگین نمونه"برابر با 1/n واریانس توزیع است
    - توزیع "میانگین نمونه"به نسبت توزیع اصلی در اطراف میانگین متمرکزتر است

$$\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

- قانون اعداد بزرگ
  - بیان می‌کند "میانگین نمونه" به میانگین توزیع نزدیک می‌شود
    - وقتی اندازه نمونه (n) بزرگ باشد

$$\lim_{n\to\infty} P\left(|\bar{X}_n - \mu| < \varepsilon\right) = 1 \quad \text{for any given number } \varepsilon > 0$$

کوواریانس و هم‌بستگی . . .

- کواریانس متغیرهای تصادفی X و Y

$$Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = Cov(Y,X)$$

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

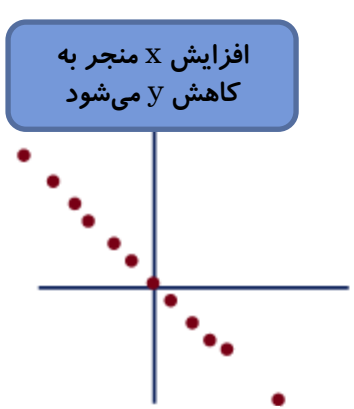
- ضریب هم‌بستگی (Correlation Coefficient)

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

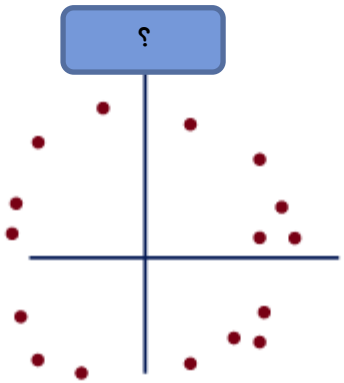
- مقدار در بازه ۱ و ۱- $-1 \le \rho(X,Y) \le 1$
- بیانگر هم‌بستگی خطی (linear dependency) بین دو متغیر X و Y
- دو متغیر تصادفی X و Y متعامد (Orthogonal) هستند اگر $E(XY) = 0$

کوواریانس و هم‌بستگی . . .

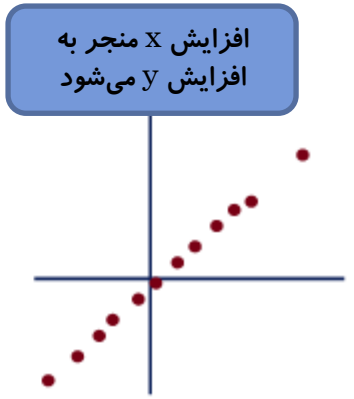
- ضریب هم‌بستگی (Correlation Coefficient)
  - اگر $\rho>0$ باشد، X و Y هم‌بستگی مثبت دارند
    - افزایش قد == افزایش وزن (کودکی تا جوانی)
  - اگر $\rho<0$ باشد، هم‌بستگی منفی دارند
    - افزایش استفاده از گوشی == کاهش باطری
  - اگر $\rho=0$ باشد، هم‌بسته (Correlated) نیستند
    - قد یک انسان با رنگ چشم وی!



ρ=-1



ρ=0



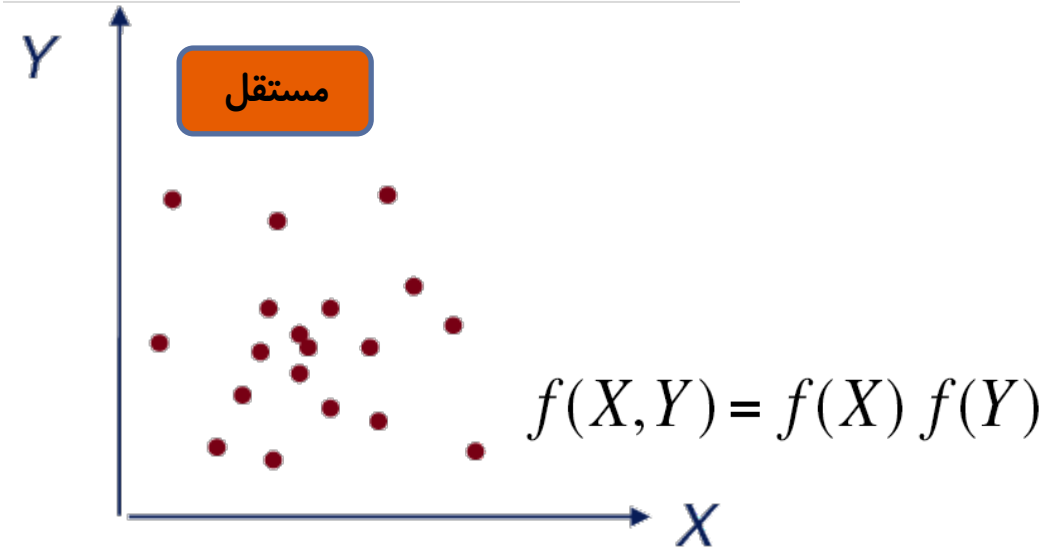
ρ=+1

کوواریانس و هم‌بستگی . . .

- **متغیرهای تصادفی مستقل (Independent)**
  - **اگر داشته باشیم**

$$Cov(X,Y) = \rho_{XY} = 0$$

- **آیا ناهم‌بسته بودن (uncorrelated)، استقلال (independence) را نتیجه می‌دهد؟**
  - مستقل بودن، ناهمبسته بودن را نیز نتیجه می‌دهد، اما برعکس آن درست نیست (به غیر از توزیع نرمال)



کوواریانس و هم‌بستگی

چند قضیه

- اگر رابطه متغیرهای X و Y به صورت $Y=aX+b$ باشد، آنگاه
  - اگر $a>0$ باشد، $\rho_{XY}=+1$
  - اگر $a<0$ باشد، $\rho_{XY}=-1$

- برای هر دو متغیر X و Y داریم

$$Var(X+Y)=Var(X)+Var(Y)+2Cov(X,Y)$$

- اگر n متغیر تصادفی $x_1, x_2, \ldots, x_n$ داشته باشیم، آنگاه

$$Var(\sum_{i=1}^{n} X_i)=\sum_{i=1}^{n} Var(X_i)+2\sum_{i=1}^{n}\sum_{j=1}^{i-1} Cov(X_i,X_j)$$

Hadi Veisi (h.veisi@ut.ac.ir)

بردارهای تصادفی . . .

بردار تصادفی

- وقتی یک متغیر تصادفی یک بردار باشد و نه یک عدد

$$\mathbf{X} = (X_1, \cdots, X_n)$$

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

برداری با n مولفه

- مشخصات دانشجویان
  - X1= وزن، X2= سن، X3= قد و ...

بردار میانگین

- یک بردار n بعدی که مؤلفه‌های آن امید ریاضی‌های تک تک مؤلفه‌های $\mathbf{X}$ است

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_n) \end{bmatrix}$$

ماتریس کوواریانس

- مؤلفه‌های قطر اصلی ماتریس کوواریانس = واریانس‌های هر کدام از X_iها
- کوواریانس متقارن است

$$Cov(X_i, X_j) = Cov(X_j, X_i)$$

$$Cov(\mathbf{X}) = E\left[[X - E(X)][X - E(X)]^T\right] = \begin{bmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_n) \\ \vdots & & \vdots \\ Cov(X_n, X_1) & \cdots & Cov(X_n, X_n) \end{bmatrix}$$

بردارهای تصادفی

- رابطه خطی بردارهای تصادفی
  - بردار n بعدی X
  - بردار m بعدی Y

یک ماتریس m×n

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{B}$$

یک بردار m بعدی

- میانگین و کواریانس Y بر حسب میانگین و کواریانس X
  - کاربرد در تبدیل ویژگی‌ها (کاهش بعد)

$$E(\mathbf{Y}) = \mathbf{A}E(\mathbf{X}) + \mathbf{B}$$

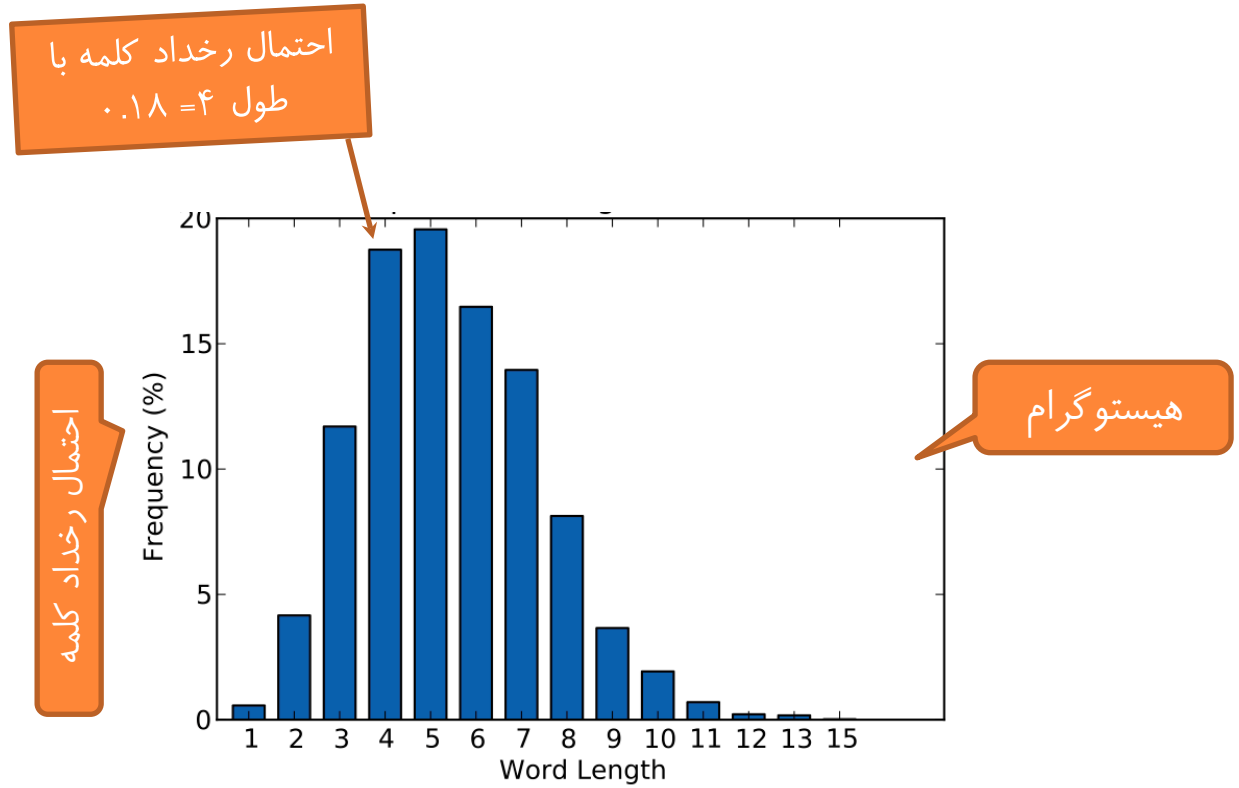
$$Cov(\mathbf{Y}) = \mathbf{A}Cov(\mathbf{X})\mathbf{A}^t$$

ترانهاده

  - کاربرد در تطبیق (adapt) پارامترهای مدل‌های آوایی و زبانی در بازشناسی گفتار

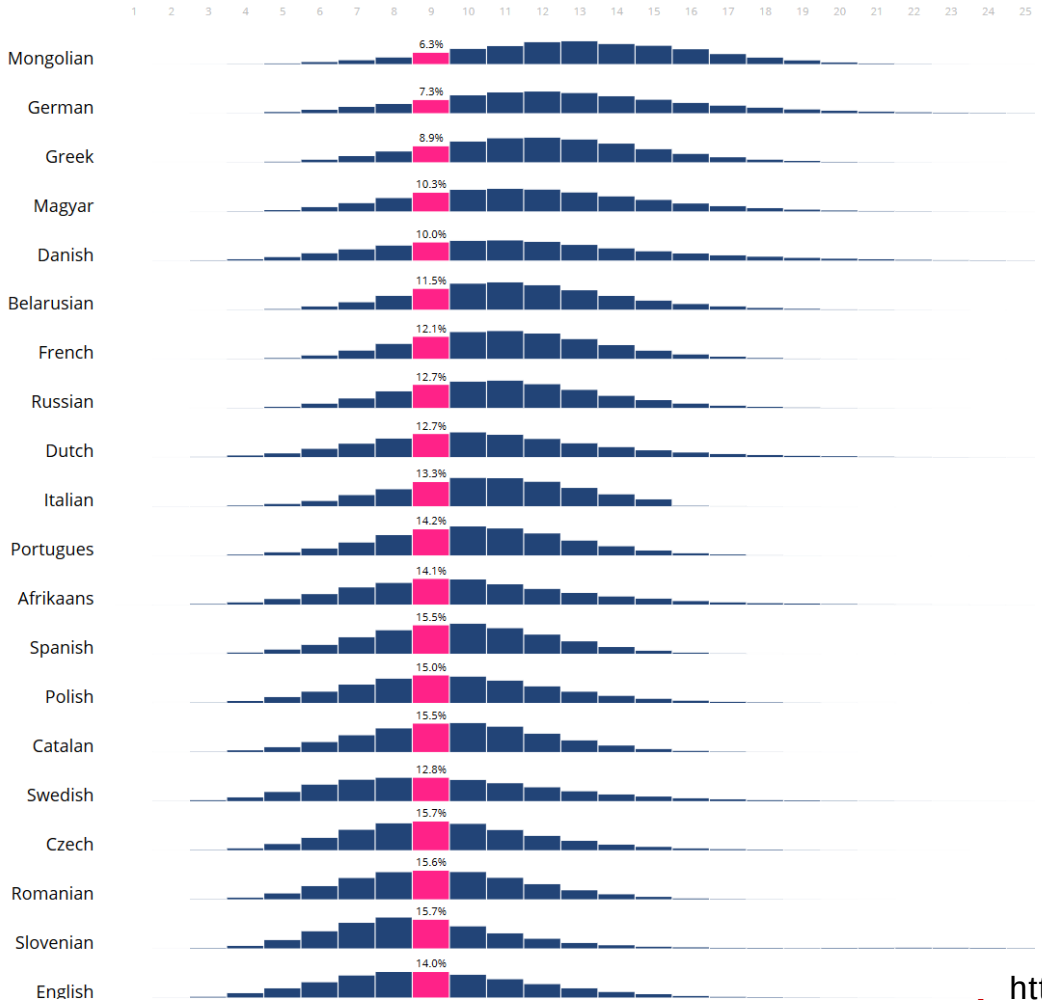
توابع توزیع . . .

- **مثال: پردازش متن (طول کلمات) . . .**
  - یک لغت‌نامه داریم
  - می‌خواهیم تعداد کلمات با طول ۱ حرف، با طول ۲ حرف، ...، طول ۲۰ حرف را بشماریم



توابع توزیع . . .

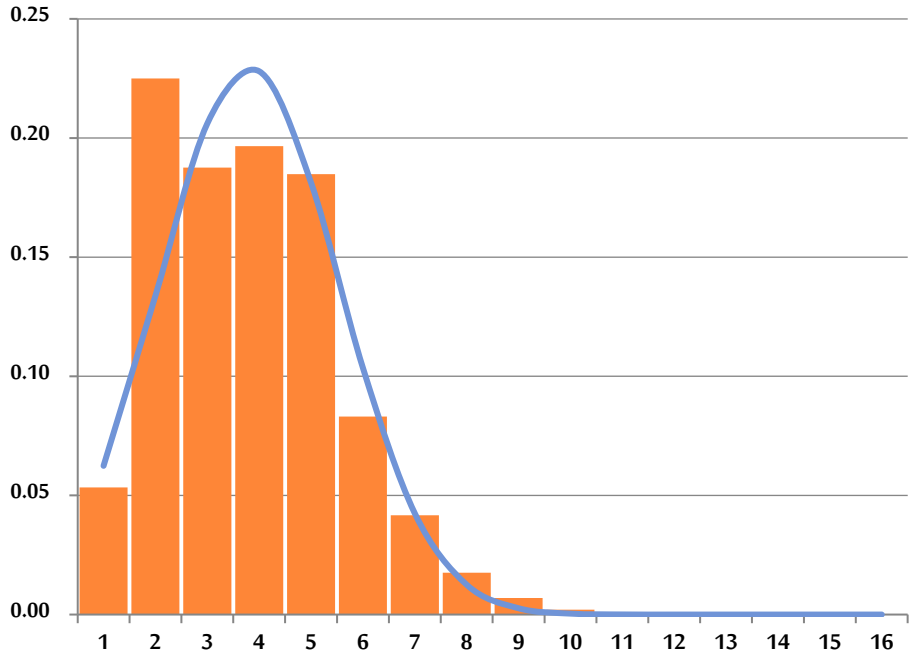
- مثال: پردازش متن (طول کلمات) . . .



http://www.ravi.io/language-word-lengths

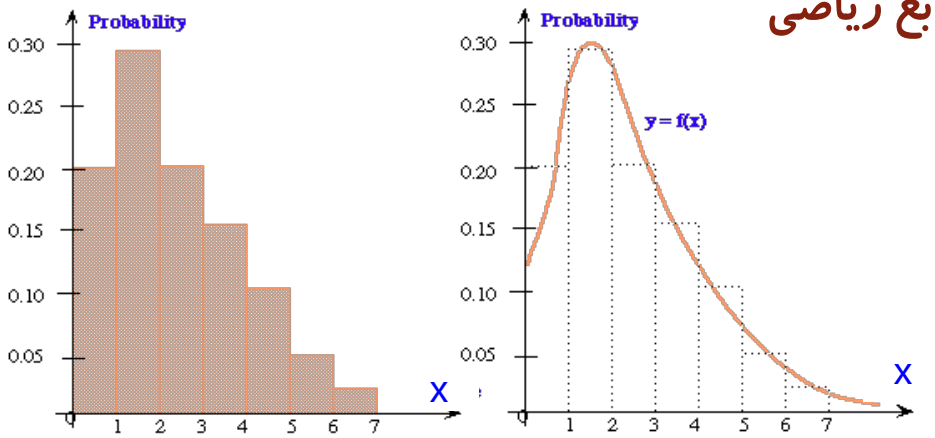
توابع توزیع . . .

- مثال: پردازش متن (طول کلمات)- برای فارسی
  - روی پیکره کوچک
    - متوسط طول کلمات: ۳.۸

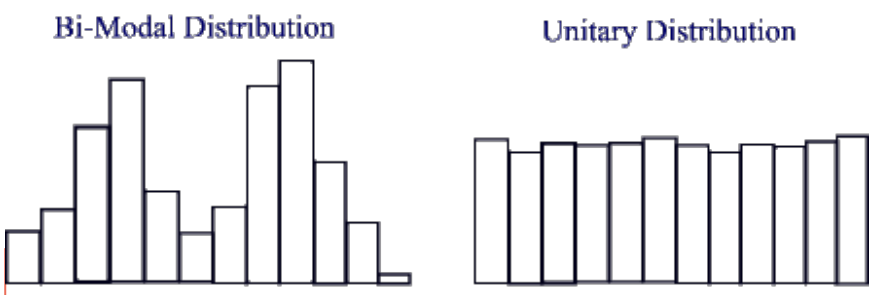


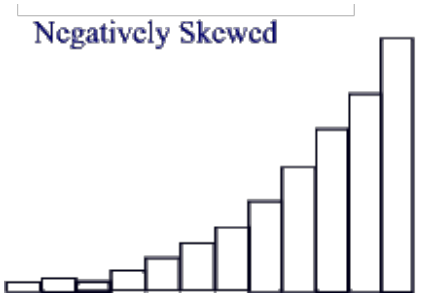
توابع توزیع . . .

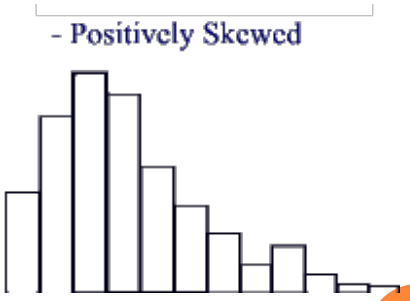
- از هیستوگرام به تابع توزیع
  - نمایش شکل توزیع احتمال‌ها با یک تابع ریاضی



  - می‌تواند شکل‌های مختلفی داشته باشد







توابع توزیع . . .

توزیع یکنواخت (Uniform Distribution)

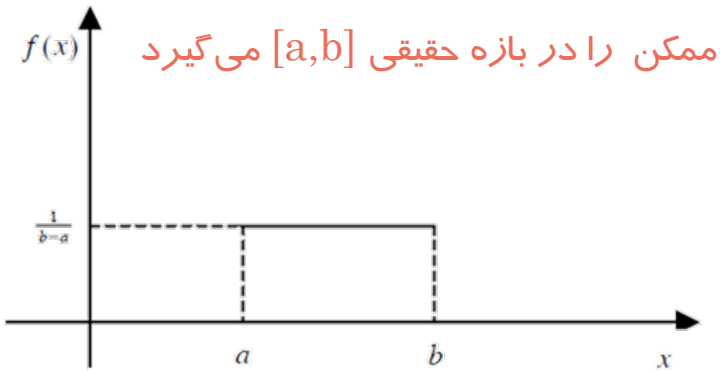
- تابع احتمال یا تابع توزیع احتمال، یک تابع ثابت است
- احتمال وقوع همه مقادیر یکسان است
  - مثال: احتمال انتخاب هر نقطه در یک بازه عددی مشخص
  - مثال: احتمال آمدن هر کدام از ۶ وجه یک تاس

- برای متغیر گسسته
  - X فقط مقادیر ممکن از ۱ تا n را می‌گیرد

$$P(X = x_i) = \frac{1}{n} \qquad 1 \le i \le n$$

- برای متغیر پیوسته
  - X فقط مقادیر ممکن را در بازه حقیقی [a,b] می‌گیرد

$$f(x) = \frac{1}{b-a} \qquad a \le x \le b$$



توابع توزیع . . .

- توزیع گاوسی (Gaussian Distribution) . . .
  - یا توزیع نرمال (Normal Distribution)
  - مهم‌ترین توزیع احتمال
    - متغیرهای تصادفی مطالعه شده در آزمایش‌های مختلف فیزیکی (از جمله سیگنال‌های گفتاری) دارای توزیع‌هایی هستند که تقریباً گاوسی است
    - محاسبات آن آسان است (به ویژه در تخمین‌ها)
    - قضیه حد مرکزی (Central Limit Theorem)

- برای یک متغیر تصادفی پیوسته

میانگین واریانس

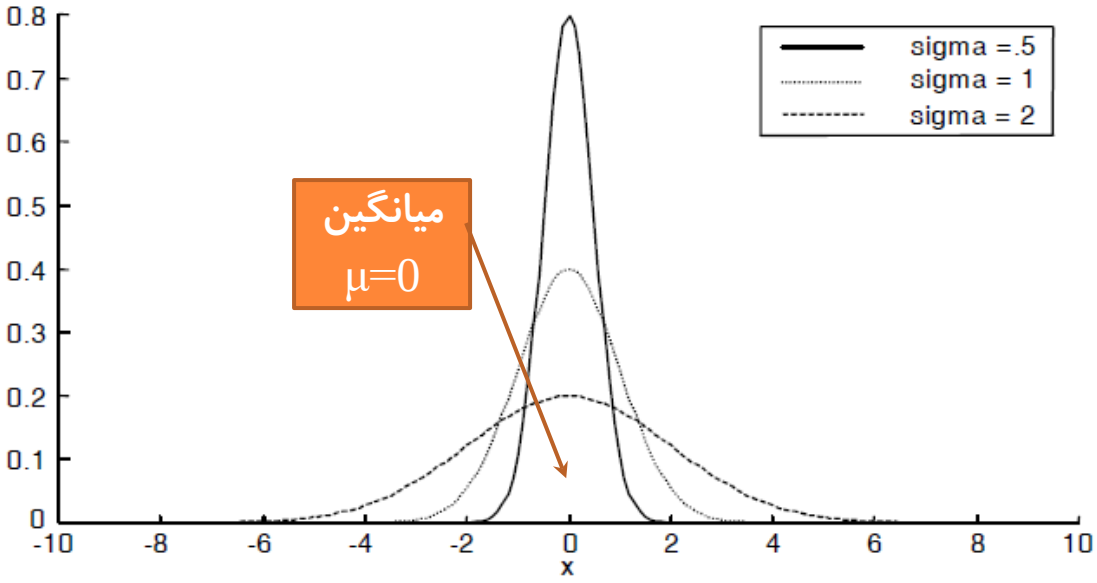
$$f(x \mid \mu, \sigma^2) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

- تابع گاوسی حول میانگین تقارنی است
- میانگین، میانه (نقطه تقارن) و نمای (بیشینه مقدار) توزیع یک نقطه است

توابع توزیع . . .

- توزیع گاوسی (Gaussian Distribution) . . .

$$f(x \mid \mu, \sigma^2) = N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$



- کاهش واریانس=تراکم بیشتر حول میانگین

توابع توزیع . . .

- **توزیع گاوسی (Gaussian Distribution) . . .**
  - **اگر متغیر تصادفی X یک توزیع گاوسی با میانگین μ و واریانس σ^2 باشد**
    آنگاه هر تابع خطی $Y=aX+b$ نیز یک توزیع گاوسی دارد
    Y یک توزیع گاوسی با میانگین $a\mu+b$ و واریانس $a^2\sigma^2$ دارد
  - **حالت کلی: مجموع $X_1+ \ldots +X_n$ از متغیرهای تصادفی مستقل $X_1, \ldots, X_n$ نیز، که در آن هر متغیر تصادفی X_i یک توزیع گاوسی دارد، یک توزیع گاوسی است**
  - **توزیع گاوسی استاندارد یا توزیع گاوسی** $N(0,1)$ =
    - میانگین صفر و واریانس یک
    - رفتار توزیع گاوسی را می‌توان فقط با استفاده از توزیع گاوسی استاندارد توضیح داد
      - تبدیل خطی توزیع گاوسی یک توزیع گاوسی است
    - اگر متغیر تصادفی X یک توزیع گاوسی با میانگین μ و واریانس σ^2 باشد، می‌توان نشان داد

$$Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

توابع توزیع . . .

توزیع گاوسی چندمتغیره (Multivariate)

- برای بردار تصادفی n بعدی

n=تعداد ابعاد بردار x(متغیرها)

$$f(\mathbf{X}=\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma}) = N(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}\,|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{t}\,\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

| . | = دترمینان

t= ترانهاده

-1 = معکوس

- میانگین

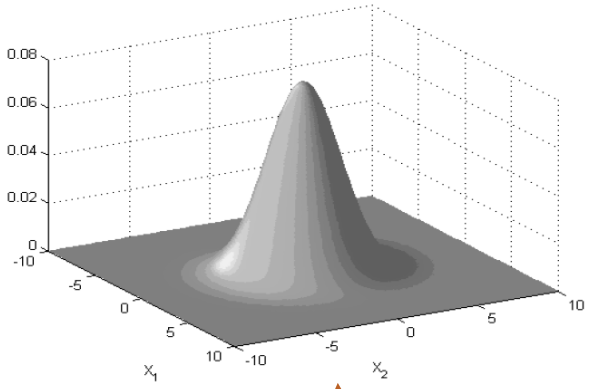
$$\boldsymbol{\mu} = E(\mathbf{x})$$

- کواریانس

$$\boldsymbol{\Sigma} = E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{t}\right]$$

  - متقارن و مثبت-معین (positive definite) (دترمینان مثبت)
  - عناصر قطر اصلی σ_{ii}= واریانس متغیر متناسب x_i
  - عناصر غیرقطر اصلی σ_{ij}= کواریانس متغیرهای x_i و x_j

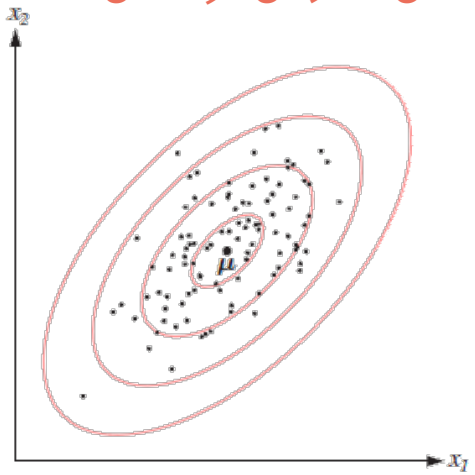
$$\sigma_{ij}^2 = E\left[(x_i-\mu_i)(x_j-\mu_j)\right]$$



برای دو بعد

توابع توزیع ...

- شکل داده‌های دارای توزیع نرمال
  - نمونه‌های داده که از توزیع نرمال پیروی می‌کنند، داخل یک خوشه قرار می‌گیرد
  - مرکز خوشه = میانگین، شکل خوشه (بیضی شکل) = تعیین شده توسط واریانس (ماتریس کواریانس)



  - بردار ویژه ماتریس کواریانس = محورهای اصلی بیضی
  - مقادیر ویژه ماتریس کواریانس = طول محورهای اصلی بیضی

- فاصله ماهالونوبیس (Mahalonobis distance)
  - فاصله بین یک مجموعه داده معین (با پارامترهای میانگین و واریانس) و یک نمونه داده
  - در نظر گرفتن وابستگی بین داده‌ها (متفاوت با فاصله اقلیدسی)
  - تغییرناپذیر با مقیاس (scale invariance)

$$(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

توابع توزیع . . .

- **قضیه حد مرکزی (Central Limit Theorem) . . .**
  - n متغیر تصادفی $X_1, \ldots, X_n$ که i.i.d هستند (مستقل و با توزیع یکسان) داریم
    - توزیع این متغیرها دارای میانگین μ و واریانس σ^2 است
  - با افزایش n به سمت بی‌نهایت، داریم
    - متغیر تصادفی Y دارای توزیع گاوسی استاندارد است

$$Y_n = \frac{n(\bar{X}_n - \mu)}{\sqrt{n\sigma^2}} \sim N(0,1)$$

میانگین نمونه‌ای متغیرها

  - متغیر تصادفی میانگین نمونه‌ای دارای توزیع گاوسی با میانگین μ و واریانس σ^2/n است

توابع توزیع . . .

- **قضیه حد مرکزی (Central Limit Theorem)**
  - **توسعه برای حالتی که توزیع‌ها یکسان نیستند** (Liapounov 1901)
    - متغیرهای تصادفی $X_1, \ldots, X_n$ مستقل هستند و $E(|X_i - \mu_i|^3) < \infty$
    - آنگاه متغیر زیر دارای توزیع گاوسی است

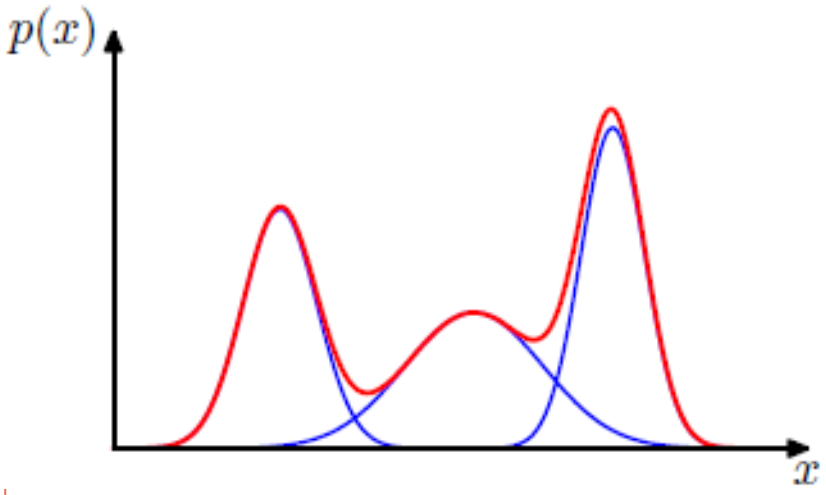
$$Y_n = \left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i\right) / \left(\sum_{i=1}^{n} \sigma_i^2\right)^{1/2}$$

- مجموع متغیرهای تصادفی $X_1, \ldots, X_n$ دارای توزیع گاوسی با میانگین $\sum_{i=1}^{n} \mu_i$ و واریانس $\left(\sum_{i=1}^{n} \sigma_i^2\right)^{1/2}$ است

حاصل جمع تعداد زیادی متغیر تصادفی مستقل، صرف‌نظر از توزیع‌های اصلی هر یک از آن‌ها، با بزرگ شدن تعداد متغیرهای تصادفی، دارای توزیع گاوسی است.

توابع توزیع

- مدل‌های مخلوط (Mixture Model)
  - ترکیب (خطی) چند مدل با همدیگر
  - مدل کردن توزیع‌های پیچیده با بیشینه‌های محلی چندگانه
  - برای توزیع نرمال (گاوسی): مدل مخلوط گاوسی (GMM: Gaussian Mixture Model)
    - از پرکاربردترین روش‌های مدل‌سازی



ضریب مخلوط

$$f(\mathbf{x}) = \sum_{k=1}^{K} c_k N_k(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$c_k \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} c_k = 1$$

نظریه اطلاعات . . .

نظریه اطلاعات

- مبنای ریاضی را برای انتقال اطلاعات
- بعد از کارهای شانون (Shannon)
  - در ۱۹۴۸

- اطلاعات قابل استخراج از پدیده مشاهده شده x_i به احتمال آن بستگی دارد.
  - اگر احتمال $P(x_i)$ کوچک باشد، حاوی اطلاعات زیادی است (پدیده بسیار نادر است)
  - اگر احتمال بزرگ باشد، حاوی اطلاعات کم است (پدیده مورد انتظار بوده و به وفور مشاهده می‌شود)

- میزان اطلاعات = عدم قطعیت

- میزان اطلاعات $I(x_i) = log \frac{1}{P(x_i)}$
  - وقتی پایه لگاریتم ۲ باشد، واحد اطلاعات بیت نامیده می‌شود

Hadi Veisi (h.veisi@ut.ac.ir)

نظریه اطلاعات: آنتروپی . . .

- آنتروپی (Entropy)
  - اگر X یک متغیر تصادفی گسسته از فضای نمونه $S = \{x_1, x_2, \dots, x_i, \dots\}$ (الفبا) باشد.
    - هر نماد x_i دارای مقدار احتمال مربوط به خود است.
    - مثال: S = مجموعه حروف زبان فارسی
  - میانگین مقدار اطلاعات (اطلاعات مورد انتظار) = آنتروپی $H(S)$

$$H(X) = E[I(X)] = \sum_S P(x_i) I(x_i) = \sum_S P(x_i) log_2 \frac{1}{P(x_i)} = E[-log_2 P(X)]$$

  - بیانگر میزان اطلاعات لازم برای مشخص کردن اینکه به طور متوسط چه نوع نمادی رخ می‌دهد = متوسط عدم قطعیت برای نماد
  - مقدار بیشینه = زمانی که احتمال همه نمادها برابر باشد (عدم قطعیت کامل)
    - توزیع یکنواخت
  - مقدار کمینه = ۰ = زمانی که احتمال یکی از نمادها برابر با یک و سایر نمادها صفر باشد
    - وجود قطعیت کامل

نظریه اطلاعات: آنتروپی . . .

- مثال
  - یک نماد (عدم وجود ابهام=قطعی) ← p(x1)=1 $H(X) = -1log_2 1 = 0$
    - عدم قطعیت = صفر
  - دو نماد با احتمال‌های برابر (بیشینه ابهام=شانس برابر انتخاب دو نماد)
    - p(x1)=p(x2)=0.5

$$H(X) = -0.5log_2 0.5 - 0.5log_2 0.5 = 1$$

  - چهار نماد با احتمال‌های برابر (بیشینه ابهام=شانس برابر انتخاب بین چهار نماد)
    - p(x1)=p(x2) =p(x3) =p(x4)=0.25

$$H(X) = -0.25log_2 0.25 - 0.25log_2 0.25 - 0.25log_2 0.25 - 0.25log_2 0.25 = 2$$

  - چهار نماد با احتمال‌های غیربرابر
    - p(x1)=p(x2) =p(x3) =0.1; p(x4)=0.7

$$H(X) = -0.1log_2 0.1 - 0.1log_2 0.1 - 0.1log_2 0.1 - 0.7log_2 0.7 = 1.36$$

نظریه اطلاعات: آنتروپی . . .

- کدگذاری اطلاعات
  - می‌خواهیم نمادهای $X=\{x1, ..xn\}$ را با دنباله‌ای از بیت‌ها (۰ و ۱) کدگذاری کنیم
  - مثال: دو نماد (مانند انداختن سکه)← ۱ بیت کافیست: شیر =۰ و خط=۱
  - مثال: چهار نماد با احتمال برابر

نظریه اطلاعات: آنتروپی . . .

- در مدل‌سازی بر مبنای انتروپی
  - مدلی بهتر است که انتروپی کمتری دارد
  - «عدم قطعیت» کمتری دارد
  - روش‌های زیادی برای کاهش انتروپی وجود دارد

نظریه اطلاعات: آنتروپی . . .

احتمال برابر برای همه حروف

- آنتروپی زبان انگلیسی (توسط شانون)
  - برای حروف (۲۶ حرف)
    - احتمال= محاسبه به صورت N-gram برای حروف
    - F_n= آنتروپی حرف nام به شرط n-1 حرف قبلی
      - در اینجا ۴.۲ بیت برای هر حرف
        - در کدگذاری ASCII هر حرف ۸ بیت دارد
    - متوسط طول هر کلمه = ۴.۵ حرف (محاسبه روی ۸۰۰۰ کلمه)
      - متوسط آنتروپی کلمه (تعداد بیت به ازای هر کلمه) = ۱۱.۸

	F_0	F_1	F_2	F_3	F_{word}
26 letter	4.70	4.14	3.56	3.3	2.62
26 letter+Space	4.76	4.03	3.32	3.1	2.14

احتمال حروف در انگلیسی

- سرگشتگی
  - تخمین میزان انشعاب (یا متوسط تعداد انتخاب‌های الفبا)
  - تقریب سرگشتگی از روی آنتروپی

$$PP(X) = 2^{H(X)}$$

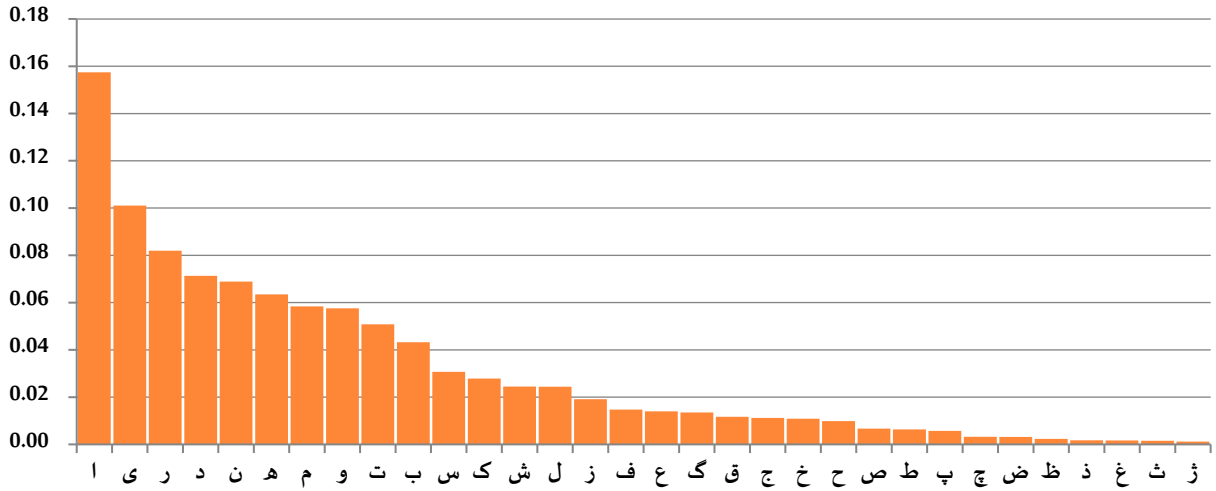
$$PP(W) = p(w_1 w_2 ... w_N)^{-\frac{1}{N}} \Rightarrow \hat{H}(W) = -\frac{1}{N}\log_2 p(w_1 w_2 ... w_N)$$

نظریه اطلاعات: آنتروپی . . .

- آنتروپی زبان فارسی
  - تعداد ۳۲ حرف

حالت محاسبه	بدون احتمال (F0)	احتمال ۱-تایی (F1)
با درنظر گرفتن فاصله	5.044	4.009
بدون درنظر گرفتن فاصله	5.000	4.249

- نمودار فراوانی نویسه‌های فارسی



نظریه اطلاعات: آنتروپی . . .

- آنتروپی متقاطع (Cross Entropy)
  - نوعی از آنتروپی برای مقایسه دو توزیع احتمال از یک پدیده
  - مقایسه میزان نزدیک بودن دو تابع توزیع احتمال
    - مقایسه احتمال تخمین زده شده و احتمال واقعی

احتمال واقعی X

$$H(X,q) = \sum_{s} p(x_i) log_2 \frac{1}{q(x_i)} = E_p[-log_2 q(X)] = H(X) + D(p||q)$$

احتمال تخمینی X (مدل p)

واگرایی (فاصله) KL: بیانگر میزان تفاوت p و q
Kullback–Leibler divergence

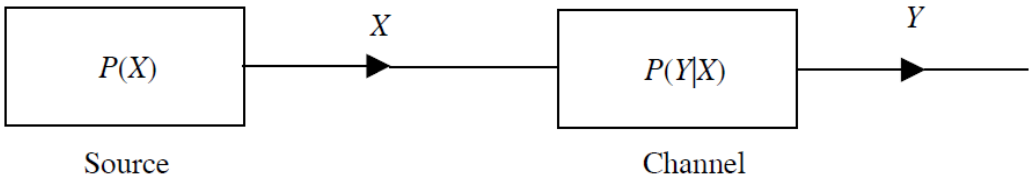
$$D(p||q) = \sum_{s} p(x_i) log_2 \frac{p(x_i)}{q(x_i)}$$

- مقدار کمینه $H(X,q)$ وقتی است که p=q باشد (مدل تخمینی q کاملا دقیق باشد)

نظریه اطلاعات: آنتروپی شرطی . . .

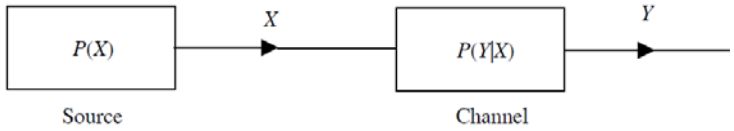
کانال اطلاعات

- هدف: انتقال نمادهای موردنظر از طریق یک کانال اطلاعاتی
- الفبای ورودی کانال: $X = (x_1, x_2, \dots, x_s)$
- الفبای خروجی کانال: $Y = (y_1, y_2, \dots, y_t)$
- کانال اطلاعاتی: نمایش با ماتریس کانال $M_{ij} = P(y_j|x_i)$
  - $P(y_j|x_i)$ احتمال شرطی دریافت نماد خروجی y_j بعد از فرستاده شدن نماد ورودی x_i



- مقدار متوسط اطلاعات (عدم قطعیت) الفبای ورودی X = آنتروپی پیشین (Prior)
  - همان $H(X)$
- بعد از انتقال، متوسط اطلاعات (عدم قطعیت) الفبای ورودی = آنتروپی پسین (Posterior)
  - آنتروپی شرطی: متوسط اطلاعات الفبای ورودی پس از انتقال به مقصد (برای فقط یک نماد)
  - $H(X|Y = y_i) = -\sum_X P(X = x_i|Y = y_j) log P(X = x_i|Y = y_j)$

نظریه اطلاعات: اطلاعات متقابل . . .



- اطلاعات متقابل (Mutual Information)

$$I(X;Y) = H(X) - H(X|Y) = \sum_X P(x_i) log \frac{1}{P(x_i)} - \sum_X \sum_Y P(x_i, y_j) log \frac{1}{P(x_i|y_j)}$$
$$= \sum_X \sum_Y P(x_i, y_j) \, log \frac{P(x_i|y_j)}{P(x_i)} = \sum_X \sum_Y P(x_i, y_j) log \frac{P(x_i, y_j)}{P(x_i)P(y_j)}$$
$$= E\left[log \frac{P(X,Y)}{P(X)P(Y)}\right]$$

- متوسط اطلاعاتی که توسط پدیده تصادفی Y درباره پدیده تصادفی X فراهم می‌شود
  - برابر است با میانگین اختلاف بین تعداد بیت‌های لازم برای مشخص کردن خروجی X، هنگامی که خروجی Y نامعلوم و هنگامی که خروجی Y معلوم است
  - بیانگر اختلاف در آنتروپی X و آنتروپی شرطی X با داشتن Y است

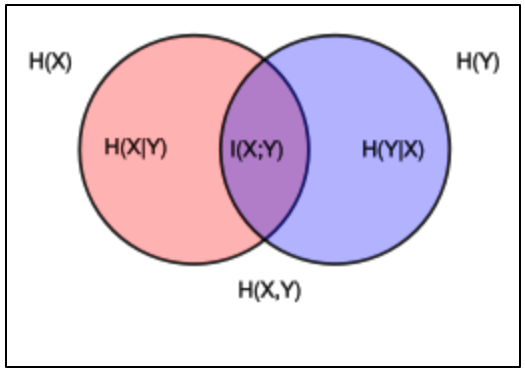
نظریه اطلاعات: اطلاعات متقابل . . .

- **بیانگر اطلاعات به‌دست‌آمده (کاهش در عدم قطعیت) از طریق یک کانال با مشاهده نماد خروجی**
  - **اگر کانال اطلاعاتی بدون نوفه باشد، با مشاهده نماد خروجی می‌توان نماد ورودی را با قطعیت تعیین کرد = آنتروپی شرطی $H(X|Y)$ برابر با صفر = کانال بدون نوفه**
    - اطلاعات متقابل بیشینه را به دست می‌آوریم $I(X;Y) = H(X)$
  - **در کاربردهای واقعی، کانال اطلاعاتی معمولاً دارای نوفه است**
    - آنتروپی شرطی $H(X|Y)$ صفر نیست
    - به حداکثر رساندن اطلاعات متقابل، معادل با به دست آوردن یک کانال اطلاعاتی با نوفه پایین است
      - که رابطه نزدیک‌تری بین نمادهای ورودی و خروجی ایجاد می‌کند

معیاری برای مدل‌سازی

نظریه اطلاعات: اطلاعات متقابل . . .

- بیانگر وابستگی متقابل دو متغیر



- استفاده به عنوان معیار فاصله

نظریه اطلاعات: اطلاعات متقابل

- **اطلاعات متقابل نقطه ای (PMI: Pointwise Mutual Information)**
  - همان اطلاعات متقابل است که برای دو نقطه (به جای دو متغیر تصادفی) تعریف می شود

- **اگر x و y هر کدام یک کلمه باشند**
  - PMI بیانگر ارتباط معنایی دو کلمه است
  - PMI('read', 'book')>PMI('read', 'dress')

  - کاربرد زیاد در پردازش متن
    - خلاصه سازی
    - هم رخدادی
    - ...

Bigram	freq1	freq2	bigram freq	PMI
great britain	2	2	2	7.22
his assent	9	4	4	7.22
britain is	2	10	2	7.06
independent states	4	8	3	6.97
united states	3	8	2	6.8
human events	1	2	1	6.22
one people	1	10	1	6.22
equal station	2	1	1	6.22
mankind requires	3	1	1	6.22
becomes necessary	1	2	1	6.22
created equal	1	2	1	6.22
.....				
and the	57	78	3	-0.14
and our	57	26	1	-0.14
and of	57	79	3	-0.16
and for	57	29	1	-0.3
of and	79	57	1	-1.75

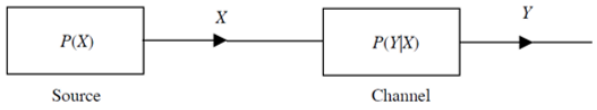
نظریه اطلاعات: کاربردها . . .

کاربردهای کانال اطلاعات در پردازش زبان طبیعی . . .

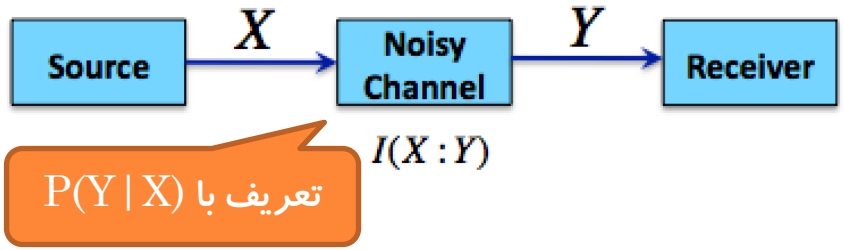
- **مدل کانال نویزی**
  - Noisy Chanel Model
- **فرآیند**
  - ارسال پیغام X از فرستنده
  - عبور پیغام از کانال (نویزی)
  - دریافت پیغام Y در گیرنده
- **در کانال غیر نویزی**
  - X و Y برابرند و $I(X;Y) = H(X)$

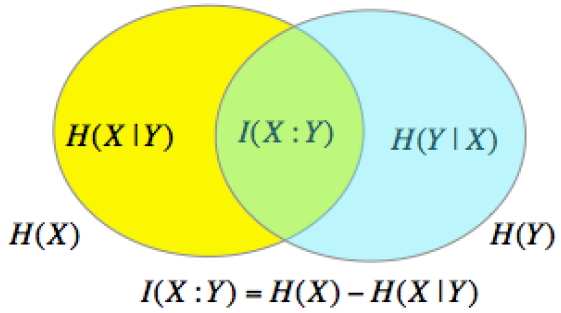
کانال اطلاعات

- هدف: انتقال نمادهای موردنظر از طریق یک کانال اطلاعاتی
- الفبای ورودی کانال: $X = (x_1, x_2, \dots, x_s)$
- الفبای خروجی کانال: $Y = (y_1, y_2, \dots, y_t)$
- کانال اطلاعاتی: نمایش با ماتریس کانال $M_{ij} = P(y_j|x_i)$
  - $P(y_j|x_i)$ احتمال شرطی دریافت نماد خروجی y_j بعد از فرستاده شدن نماد ورودی x_i



- مقدار متوسط اطلاعات (عدم قطعیت) الفبای ورودی X = آنتروپی پیشین (Prior)
  - همان $H(X)$
- بعد از انتقال، متوسط اطلاعات (عدم قطعیت) الفبای ورودی = آنتروپی پسین (Posterior)





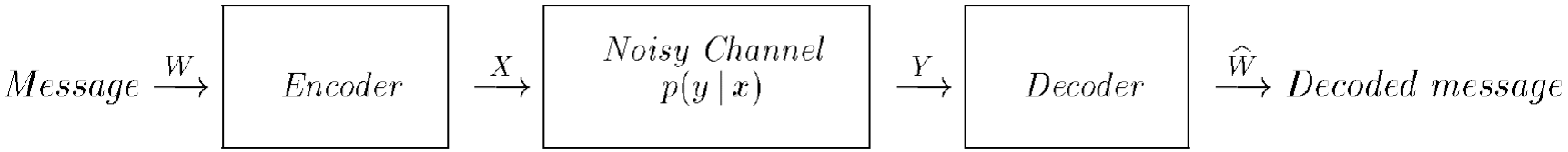
نظریه اطلاعات: کاربردها . . .

- کاربرد مدل کانال نویزی در خطایاب املایی
  - اطلاعات فرستند (X): متن واقعی
    - X= مجموعه‌ای از حروف (فارسی/انگلیسی)
  - اطلاعات گیرنده (Y): متن تایپ شده
    - Y= مجموعه‌ای از حروف (فارسی/انگلیسی)
  - کانال نویزی: اشتباه‌های تایپی در متن تایپ شده
    - $P(Y \mid X)$ = ماتریس کانال = احتمال اینکه بخواهیم X بنویسیم ولی به جای آن Y نوشته شده باشد
      - ماتریسی است که به ازای هر حرف، احتمال جایگزینی/حذف/درج آن با سایر حروف را تعیین می‌کند.

نظریه اطلاعات: کاربردها . . .

کاربرد مدل کانال نویزی در مترجم ماشینی

- حالت کلی‌تر مدل



- مثال: ترجمه از فارسی به انگلیسی
- اطلاعات فرستند (X): متن زبان مقصد (انگلیسی)
  - X= دنباله کلمات زبان مقصد (انگلیسی)
- اطلاعات گیرنده (Y): متن زبان مبدا (فارسی)
  - Y= دنباله‌ای از کلمات زبان مبدا (فارسی)

$$\hat{W} = \underset{x}{\operatorname{argmax}} P(x|y)$$
$$= \underset{x}{\operatorname{argmax}} \frac{P(x)P(y|x)}{P(y)}$$
$$= \underset{x}{\operatorname{argmax}} P(x)P(y|x)$$

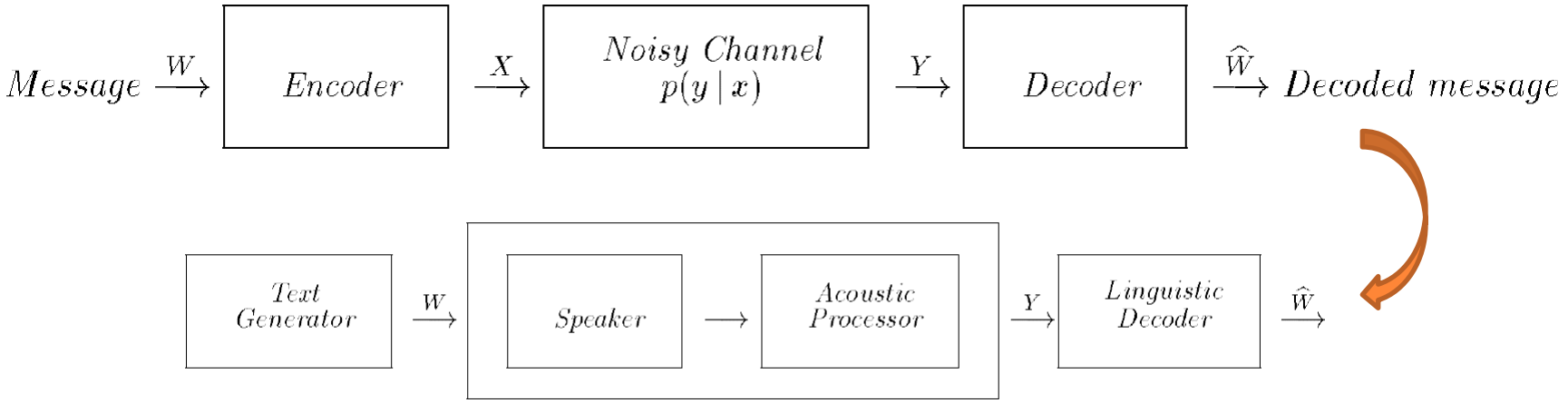
مدل زبانی

مدل کانال= مدل ترجمه

- کانال نویزی
  - $P(Y \mid X)$ = احتمال اینکه دنباله Y دارای ترجمه X باشد

نظریه اطلاعات: کاربردها . . .

کاربرد مدل کانال نویزی در بازشناسی گفتار



- اطلاعات فرستند (X): کلمات
- اطلاعات گیرنده (Y): سیگنال
- کانال
  - P(Y | X) = احتمال اینکه سیگنال Y معادل متن X باشد

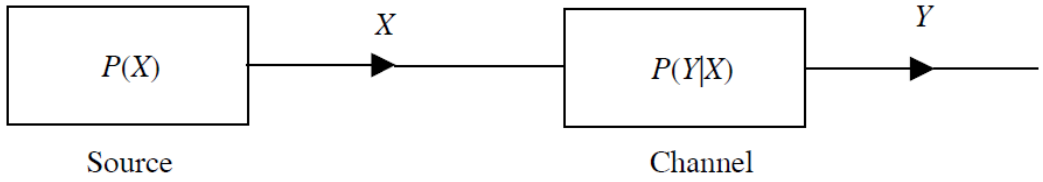
$$\hat{W} = \underset{x}{\operatorname{argmax}} P(x|y)$$
$$= \underset{x}{\operatorname{argmax}} \frac{P(x)P(y|x)}{P(y)}$$
$$= \underset{x}{\operatorname{argmax}} P(x)P(y|x)$$

مدل زبانی

مدل کانال= مدل آوایی

نظریه اطلاعات: کاربردها

- **کاربرد مدل کانال نویزی در پردازش زبان و گفتار**

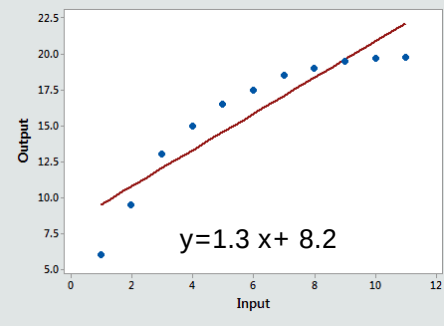


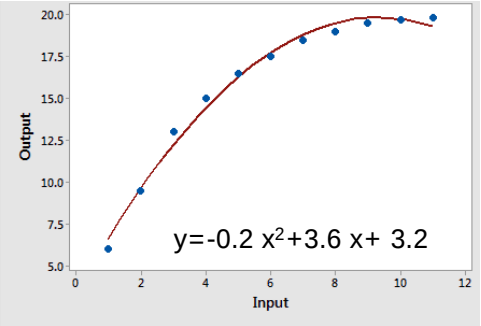
کاربرد	ورودی (X)	خروجی (Y)	P(X)	P(Y | X)
ترجمه ماشینی	دنباله کلمات زبان مقصد(L1)	دنباله کلمات زبان مبدا (L1)	مدل زبانی زبان مقصد= P(L1)	مدل ترجمه
غلط یاب املایی	متن واقعی	متن دارای خطای املایی	احتمال رخداد متن‌ها	مدل خطاهای املایی
برچسپ گذاری پاره گفتار	دنباله برچسپ‌ها	دنباله کلمات	احتمال دنباله برچسپ‌ها	احتمال کلمه به شرط برچسپ
بازشناسی گفتار	دنباله کلمات	دنباله سیگنال گفتار	احتمال دنباله کلمات (مدل زبانی)	مدل آوایی

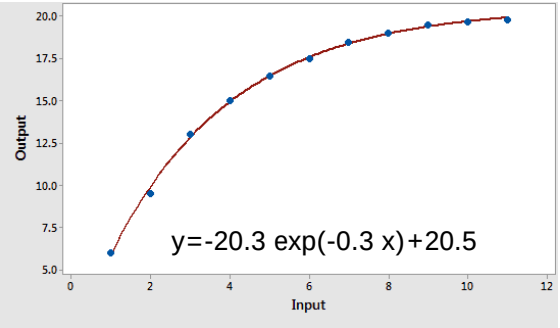
نظریه تخمین . . .

مساله

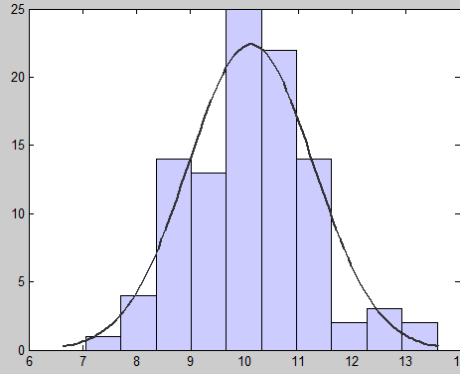
- تعدادی نمونه داده داریم، می‌خواهیم آنها را با یک تابع (خطی/غیرخطی) مدل کنیم
  - مثال: وزن افراد بر حسب قد آنها، سیگنال تمیز بر حسب سیگنال نویزی







- تعدادی نمونه داده داریم، می‌خواهیم تابع توزیع احتمال آنها را بدست آوریم
  - مثال: توزیع گاوسی به طول کلمات در یک زبان



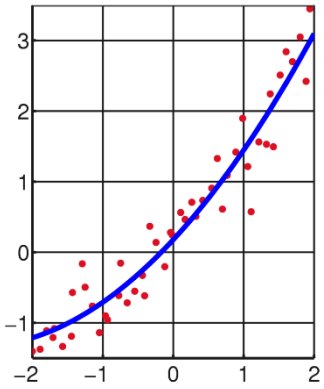
$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

نظریه تخمین . . .

- نظریه تخمین (Estimation theory)
  - در مدل‌سازی آماری یک تابع توزیع (مانند گاوسی) برای داده‌ها فرض می‌کنیم
  - و باید از روی داده‌های آموزش، پارامترهای آن توزیع (مانند میانگین و واریانس) را تخمین بزنیم

  - متغیرهای تصادفی $X_1, \ldots, X_n$ را که i.i.d هستند (مستقل و با توزیع یکسان) داریم
  - هدف تخمین پارامترهای Φ
  - تابع تخمین‌گر $\theta(X_1, \ldots, X_n)$

- روش‌های تخمین
  - کمینه میانگین مربعات خطا (MMSE: Minimum Mean Square Error)
  - تخمین بیشینه شباهت (MLE: Maximum-Likelihood Estimation)
  - تخمین بیز (Bayesian Estimation)



نظریه تخمین . . .

کمینه میانگین مربعات خطا (MMSE) . . .

- کمینه کردن امید ریاضی مربعات خطای بین مقدار واقعی و مقدار تخمین زده شده

$$E(Y-\hat{Y})^2 = E(Y-g(X))^2$$

- فرض کنید هدف ما تخمین مقدار Y با داشتن X باشد، یعنی $\hat{Y} = g(X)$
- که $g(X)$ تابعی بر حسب پارامترهای Φ است، یعنی $g(X,\Phi)$
  - و با داشتن پارامترهای Φ، تابع $g()$ به صورت کامل مشخص می‌شود
  - پس: هدف تخمین پارامترهای Φ است

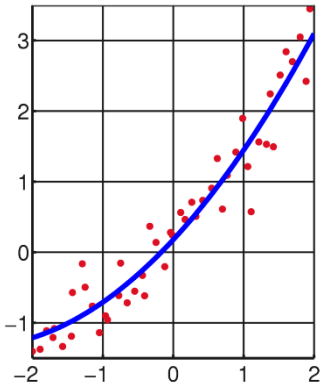
$$\hat{\mathbf{\Phi}}_{MMSE} = \arg\min_{\Phi}\left[E\left[(Y-g(X,\mathbf{\Phi}))^2\right]\right]$$

- تخمین LSE: Least Square Error
  - در عمل به جای تابع توزیع توام X و Y، نمونه‌هایی از x_i و y_i معادل داریم

$$\mathbf{\Phi}_{LSE} = \arg\min_{\Phi}\sum_{i=1}^{n}\left[y_i - g(x_i,\mathbf{\Phi})\right]^2$$

  - قانون اعداد بزرگ: وقتی تعداد نمونه‌ها به بی‌نهایت میل می‌کند، LSE و MMSE برابر می‌شوند

نظریه تخمین . . .



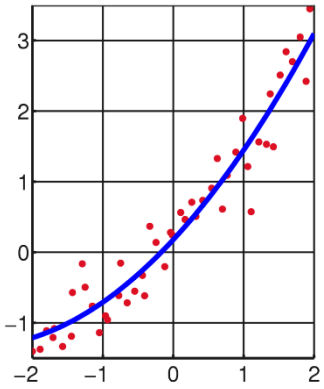
- کمینه میانگین مربعات خطا (MMSE): برای تابع ثابت . . .
  - تابع ثابت $\hat{Y} = g(x) = c$
    - پارامتر = c
  - هدف کمینه کردن خطاست $E(Y-\hat{Y})^2 = E(Y-c)^2$
    - مشتق گرفتن و برابر صفر قرار دادن $c_{MMSE} = E(Y)$
  - خطای مجذور میانگین کمینه دقیقاً برابر با واریانس Y است

- تخمین LSE
  - میانگین نمونه‌ای

$$\min \sum_{i=1}^{n} [y_i - c]^2$$

$$c_{LSE} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

نظریه تخمین . . .



- کمینه میانگین مربعات خطا (MMSE): برای تابع خطی . . .
  - تابع خطی $\hat{Y} = g(x) = ax + b$
    - پارامترها: a و b

$$e(a,b) = E(Y-\hat{Y})^2 = E(Y-ax-b)^2$$

$$\frac{\partial e}{\partial a} = 0, \quad \text{and} \quad \frac{\partial e}{\partial b} = 0$$

$$a = \frac{\text{cov}(X,Y)}{Var(X)} = \rho_{XY}\frac{\sigma_Y}{\sigma_X}$$

$$b = E(Y) - \rho_{XY}\frac{\sigma_Y}{\sigma_X}E(X)$$

- برای تخمین LSE
  - فرض: بردار x دارای d بعد است و n نمونه داریم

$$\hat{\mathbf{Y}} = \mathbf{XA} \text{ or } \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & \cdots & x_1^d \\ 1 & x_2^1 & \cdots & x_2^d \\ \vdots & \vdots & & \vdots \\ 1 & x_n^1 & \cdots & x_n^d \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix}$$

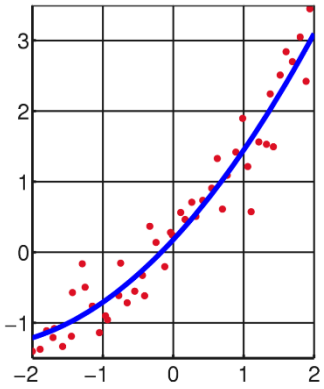
$$e(\mathbf{A}) = \| \hat{\mathbf{Y}} - \mathbf{Y} \|^2 = \sum_{i=1}^{n} \left(\mathbf{A}^t \mathbf{x}_i - y_i \right)^2$$

$$\nabla e(\mathbf{A}) = \sum_{i=1}^{n} 2(\mathbf{A}^t \mathbf{x}_i - y_i)\mathbf{x}_i = 2\mathbf{X}^t(\mathbf{XA} - \mathbf{Y})$$

$$\mathbf{X}^t\mathbf{XA} = \mathbf{X}^t\mathbf{Y}$$

$$\mathbf{A}_{LSE} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$

شبه معکوس



نظریه تخمین . . .

- کمینه میانگین مربعات خطا (MMSE): برای تابع غیرخطی
  - تابع غیرخطی

$$\min_{g(\bullet)\in G_{nl}} E\left[Y - g(X)\right]^2$$

$$\hat{Y} = g_{MMSE}(X) = E_{Y|X}(Y \mid X) = \int_{-\infty}^{\infty} y f_Y(y \mid X = x)dy$$

  - در بهسازی گفتار

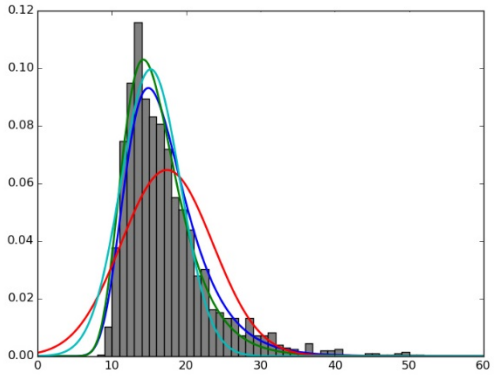
تخمین طیف گفتار تمیز

$$\hat{S}(\omega_k) = E[S(\omega_k)|\ ] = \frac{P_{ys}(\omega_k)}{P_{yy}(\omega_k)} = \frac{P_{ss}(\omega_k)}{P_{ss}(\omega_k) + P_{dd}(\omega_k)} Y(\omega_k)$$

سیگنال گفتار نویزی

توان طیف گفتار تمیز

توان طیف نویز



نظریه تخمین . . .

- تخمین بیشینه شباهت (MLE) . . .
  - پرکاربردترین روش تخمین پارامتری

- تخمین توزیع n نمونه داده i.i.d به صورت $p(x \mid \Phi) = X_1, \ldots, X_n$
- فرض: پارامترهای Φ دارای مقادیر ثابت، اما نامشخص هستند
  - اگر تابع توزیع گاوسی باشد، $\Phi=\{\mu, \Sigma\}$
- تخمین پارامترهای توزیع به نحوی که احتمال بدست آوردن نمونه داده‌ها از روی این توزیع بیشینه باشد
- چون متغیرهای تصادفی مستقل هستند
- هدف: بیشینه کردن تابع درست‌نمایی

تابع درست‌نمایی

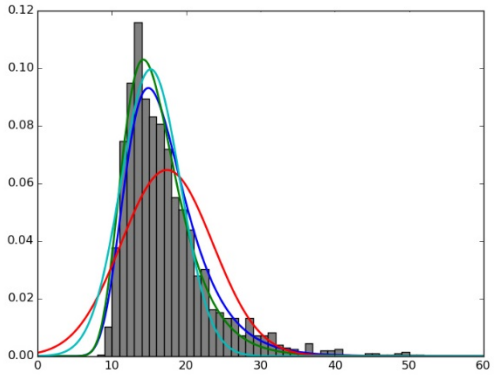
$$p_n(\mathbf{x} \mid \mathbf{\Phi}) = \prod_{k=1}^{n} p(x_k \mid \mathbf{\Phi})$$

$$\mathbf{\Phi}_{MLE} = \underset{\Phi}{\operatorname{argmax}}\, p_n(\mathbf{x} \mid \mathbf{\Phi})$$

- لگاریتم شباهت (Log-Likelihood)
  - عدم تغییر مساله (تابع یکنوای صعودی)
  - ساده کردن محاسبات و فرمول‌ها (تبدیل ضرب به جمع)

$$l(\mathbf{\Phi}) = \log p_n(\mathbf{x} \mid \mathbf{\Phi}) = \sum_{k=1}^{n} \log p(x_k \mid \mathbf{\Phi})$$

نظریه تخمین . . .



- تخمین بیشینه شباهت (MLE) . . .
  - بیشینه کردن تابع درست‌نمایی (یا لگاریتم آن) با گرادیان
    - مشتق گرفتن برحسب پارامترها و برابر صفر قرار دادن

$$\nabla_{\Phi} = \begin{bmatrix} \frac{\partial}{\partial \Phi_1} \\ \vdots \\ \frac{\partial}{\partial \Phi_k} \end{bmatrix}$$

$$\nabla_{\Phi}\, l(\boldsymbol{\Phi}) = \sum_{k=1}^{n} \nabla_{\Phi} \log p(x_k \mid \boldsymbol{\Phi}) = 0$$

- مثال ۱: توزیع گاوسی تک متغیره
  - لگاریتم تابع درست‌نمایی

$$p(x \mid \boldsymbol{\Phi}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right]$$

$$\log p_n(\mathbf{x} \mid \boldsymbol{\Phi}) = \sum_{k=1}^{n} \log p(x_k \mid \boldsymbol{\Phi}) = \sum_{k=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_k-\mu)^2}{2\sigma^2} \right] \right) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2$$

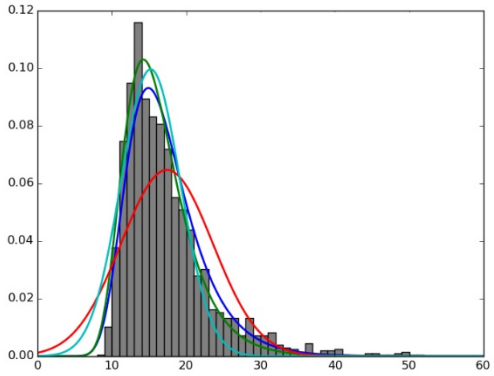
$$\frac{\partial}{\partial \mu} \log p_n(x \mid \boldsymbol{\Phi}) = \sum_{k=1}^{n} \frac{1}{\sigma^2}(x_k - \mu) = 0$$

$$\frac{\partial}{\partial \sigma^2} \log p_n(x \mid \boldsymbol{\Phi}) = -\frac{n}{2\sigma^2} + \sum_{k=1}^{n} \frac{(x_k-\mu)^2}{2\sigma^4} = 0$$

$$\mu_{MLE} = \frac{1}{n}\sum_{k=1}^{n} x_k = E(x)$$

$$\sigma^2_{MLE} = \frac{1}{n}\sum_{k=1}^{n} (x_k - \mu_{MLE})^2 = E\left[(x - \mu_{MLE})^2 \right]$$

همان میانگین و واریانس نمونه‌ای



نظریه تخمین . . .

تخمین بیشینه شباهت (MLE)

- مثال ۲: توزیع گاوسی چندمتغیره

$$p(\mathbf{x}\,|\,\boldsymbol{\Phi}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\hat{\boldsymbol{\mu}}_{MLE} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k$$

$$\hat{\boldsymbol{\Sigma}}_{MLE} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\boldsymbol{\mu}}_{MLE})(\mathbf{x}_k-\hat{\boldsymbol{\mu}}_{MLE})^t = E\left[(\mathbf{x}_k-\hat{\boldsymbol{\mu}}_{MLE})(\mathbf{x}_k-\hat{\boldsymbol{\mu}}_{MLE})^t\right]$$

- تخمین ML برای واریانس بایاس شده است

$$\mathcal{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2\right] = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

  - امید ریاضی واریانس تخمینی با واریانس واقعی برابر نیست
    - با میل کردن n به سمت بی نهایت اثر بایاس کم می‌شود

- تخمین غیربایاس

$$\hat{\sigma}^2_{\text{unbiased}} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\hat{\mu})^2$$

نظریه تخمین . . .

- **تخمین بیز (Bayesian Estimation) . . .**
  - پارامتر Φ یک متغیر تصادفی و نامشخص است
    - در تخمین ML این پارامتر تصادفی نیست و ثابت است
    - تصادفی بودن پارامتر Φ به معنی وجود احتمال پیشین توزیع $p(\Phi)$ برای آن است
  - شکل تابع توزیع $p(x \mid \Phi)$ مشخص است (مثلاً توزیع نرمال)
  - مجموعه n نمونه داده i.i.d $D=\{x_1, x_2, \ldots, x_n\}$ دارای توزیع $p(x \mid \Phi)$
    - داده‌ها حاوی اطلاعاتی از پارامتر Φ

احتمال پسین: احتمال پارامترپس از مشاهده داده‌ها

- با توجه به قانون بیز

$$p(\Phi \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \Phi) p(\Phi)}{p(\mathbf{x})} \propto p(\mathbf{x} \mid \Phi) p(\Phi)$$

$$posterior = \frac{likelihood \times prior}{evidence}$$

- هدف

$$p(x \mid D) = \int p(x \mid \Phi) p(\Phi \mid D) d\Phi$$

نظریه تخمین . . .

تخمین بیز (Bayesian Estimation) . . .

- نخستین گام در تخمین بیز: محاسبه $p(\Phi \mid \mathcal{D})$

$$p(\Phi \mid D) = \frac{1}{\alpha} p(D \mid \Phi) p(\Phi)$$

  - که α ثابت نرمال کننده است (مستقل از Φ)
    - حذف نمی‌شود
    - در عمل می‌توان مقدار ثابتی را به عنوان تخمین آن قرار داد

$$p(x \mid D) = \int p(x \mid \Phi) p(\Phi \mid D) d\Phi$$

- با فرض مستقل بودن نمونه داده‌ها

$$p(D \mid \Phi) = \prod_{k=1}^{n} p(x_k \mid \Phi)$$

نظریه تخمین . . .

- تخمین بیشینه احتمال پسین (MAP: maximum a posteriori) . . .
  - هدف: بیشینه کردن $p(\Phi \mid x)$
    - در ML هدف بیشینه کردن $p(x \mid \Phi)$
    - پارامترها متغیرهای تصادفی با توزیع پیشین $p(\Phi)$ هستند
    - متداول‌ترین تخمین‌گر بیزی

$$p(\Phi \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \Phi) p(\Phi)}{p(\mathbf{x})} \propto p(\mathbf{x} \mid \Phi) p(\Phi)$$

$$\Phi_{MAP} = \theta_{MAP}(\mathbf{x}) = \underset{\Phi}{\operatorname{argmax}}\ p(\Phi \mid \mathbf{x}) = \underset{\Phi}{\operatorname{argmax}}\ p(\mathbf{x} \mid \Phi) p(\Phi)$$

$$\Phi_{MAP} = \underset{\Phi}{\operatorname{argmax}}\ \log p(\mathbf{x} \mid \Phi) + \log p(\Phi)$$

$$\frac{\partial \log p(\mathbf{x} \mid \Phi)}{\partial \Phi} + \frac{\partial \log p(\Phi)}{\partial \Phi} = 0$$

$$\left.\frac{\partial \log p(\mathbf{x} \mid \Phi)}{\partial \Phi}\right|_{\Phi=\Phi_{MAP}} = \left.\frac{-\partial \log p(\Phi)}{\partial \Phi}\right|_{\Phi=\Phi_{MAP}}$$

- تخمین ML و MAP یکسان هستند وقتی توزیع پیشین $p(\Phi)$ یکنواخت باشد

نظریه تخمین

- **تخمین بیشینه احتمال پسین (MAP: maximum a posteriori)**
  - **مثال (قبلی): داده‌ها با توزیع گاوسی، تک متغیره، واریانس معلوم σ^2 و میانگین نامعلوم Φ**
    - توزیع پیشین Φ: گاوسی با با میانگین μ و واریانس v^2

$$p(\Phi \mid \mathbf{x}) = \frac{1}{\sqrt{2\pi}\tau} \exp\left[\frac{-1}{2\tau^2}(\Phi - \rho)^2\right]$$

    - مشتق‌گیری نسبت به Φ
    - متوسط وزن‌دار میانگین نمونه‌ها و میانگین قبلی

$$\Phi_{MAP} = \rho = \frac{\sigma^2 \mu + n v^2 \bar{x}_n}{\sigma^2 + n v^2}$$

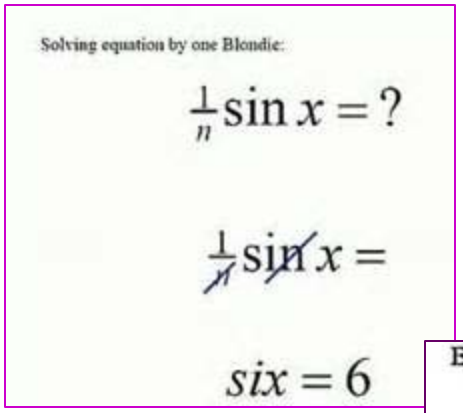
میانگین نمونه‌ها

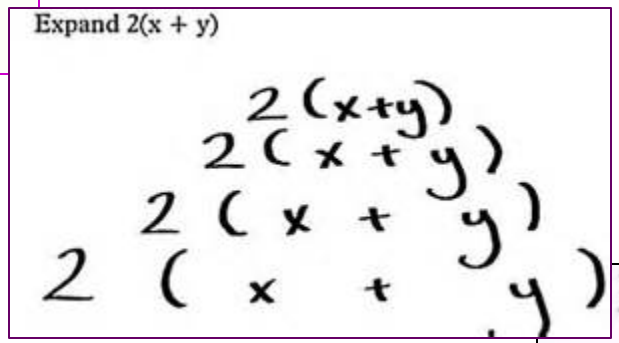
تعداد نمونه‌ها

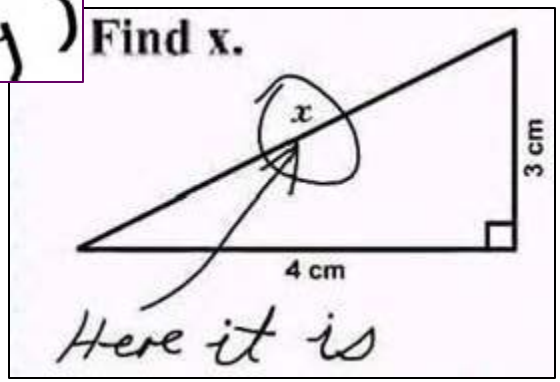
- **کاربرد در تطبیق (adaptation)**
  - آموزش صدای یک کاربر جدید به سیستم بازشناسی گفتار
  - آموزش پارامترهای مدل با پایگاه داده‌های مستقل از گوینده (با چندین گوینده) = توزیع پیشین
  - تطبیق با محاسبه میانگین نمونه‌های یک گوینده خاص و استفاده از رابطه بالا

مقایسه تخمین‌گرها

تخمین‌گر بیز	تخمین‌گر بیشینه شباهت (ML)	تخمین‌گر بیشینه احتمال پسین (MAP)
تخمین‌گر توزیع	تخمین‌گر نقطه	تخمین‌گر نقطه
$p(\mathbf{x}\|\mathcal{D}) = \int p(\mathbf{x}\|\theta)p(\theta\|\mathcal{D})d\theta$	$p(\mathbf{x}\|\mathcal{D}) = p(\mathbf{x}\|\hat{\theta})$	$p(\mathbf{x}\|\mathcal{D}) = p(\mathbf{x}\|\hat{\theta})$
تخمین توزیع	تخمین نقطه	تخمین نقطه
$p(\theta\|\mathcal{D}) = \frac{1}{\alpha}p(\mathcal{D}\|\theta)p(\theta)$	$\hat{\theta} = \arg\max_{\theta} \ln p(\mathcal{D}\|\theta)$	$\hat{\theta} = \arg\max_{\theta} \ln p(\mathcal{D}\|\theta)p(\theta)$
1. استفاده بیشتر از اطلاعات 2. کارایی بهتر در صورت عدم سازگاری بین توزیع فرض شده و توزیع واقعی 3. در نظر گرفتن بایاس واریانس	1. تفسیر ساده‌تر (نقطه‌ای) 2. محاسبات کم‌تر	1. استفاده از اطلاعات پیشین پارامتر توزیع


Solving equation by one Blondie:
$\frac{1}{n}\sin x = ?$
$\frac{1}{n}\sin x =$
$six = 6$


Expand 2(x + y)
2(x+y)
2(x+y)
2(x+y)
2(x+y)


Find x.
x
3 cm
4 cm
Here it is