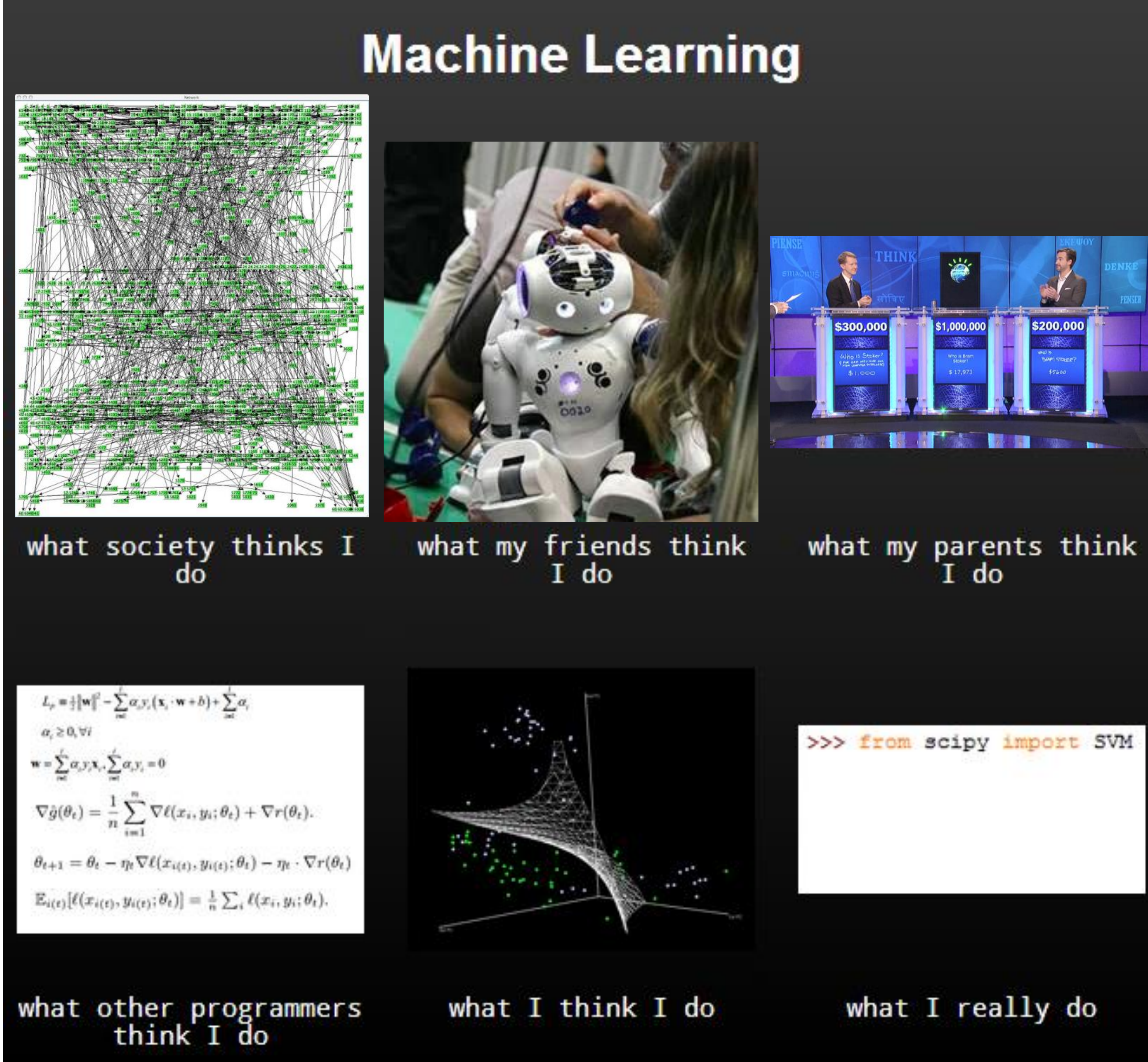
CSE 575: Statistical Machine Learning

Jingrui He

CIDSE, ASU

What is ***Machine Learning***?


Machine Learning
$300,000
$1,000,000
$200,000
what society thinks I do
what my friends think I do
what my parents think I do
>>> from scipy import SVM
what other programmers think I do
what I think I do
what I really do

Machine Learning

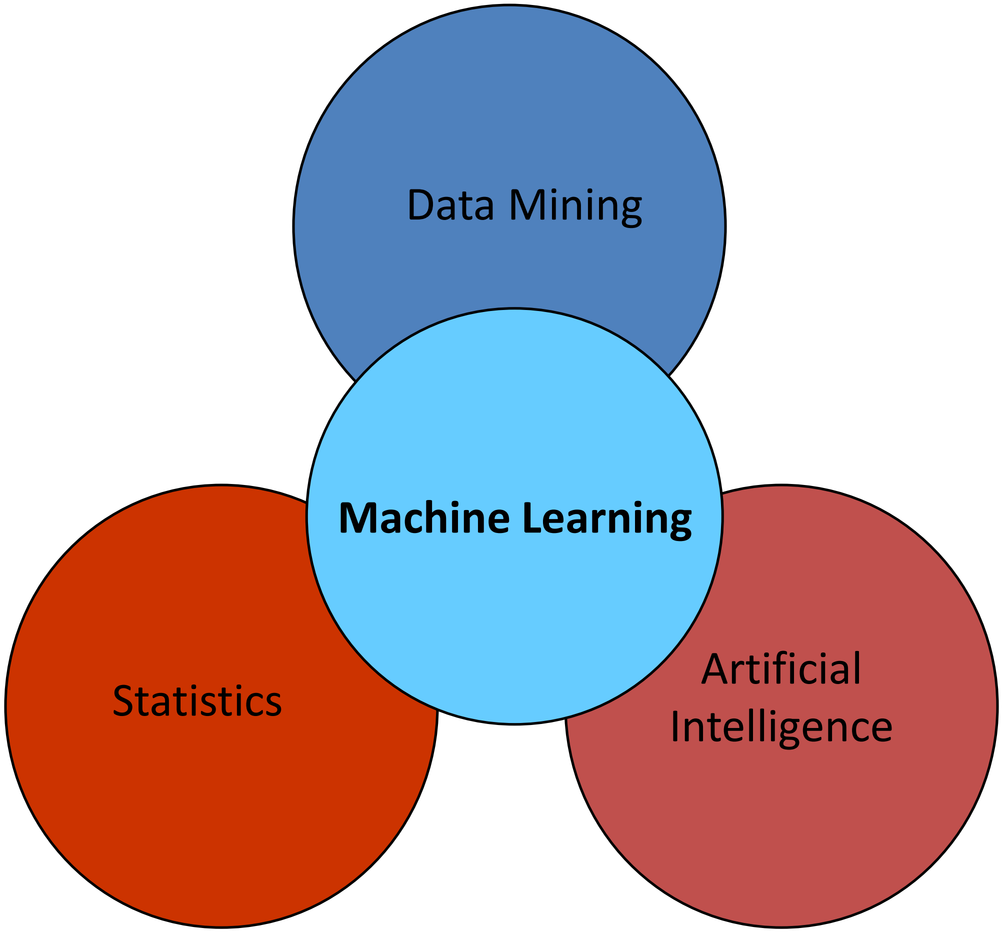
- Prof. Tom Mitchell@CMU &
- Prof. Carlos Guestrin@UW

  ‘Study of algorithms that improve their performance, at some task, with experience’

- Prof. Andrew Ng@Stanford

  ‘Machine learning is the science of getting computers to act without being explicitly programmed’

Related Fields



Useful Resources

- The discipline of machine learning: http://www.cs.cmu.edu/~tom/pubs/MachineLearning.pdf
- Coursera: https://www.coursera.org/course/ml
- Andrew Moore's tutorials: http://www.autonlab.org/tutorials/
- Alex Smola@CMU's machine learning lectures: https://www.youtube.com/playlist?list=PLZSO_6-bSqHQmMKwWVvYwKreGu4b4kMU9
- Mathworks Matlab tutorials: http://www.mathworks.com/academia/student_center/tutorials/launchpad.html
- Ben Taskar@UW's Matlab tutorial: https://alliance.seas.upenn.edu/~cis520/wiki/index.php?n=Recitations.MatlabTutorial
- Probability review by David Blei@Columbia: http://www.cs.princeton.edu/courses/archive/spring07/cos424/scribe_notes/0208.pdf

Becoming Famous in ML?

- Winning Netflix prize?
  http://en.wikipedia.org/wiki/Netflix_Prize
- Building Watson to win Jeopardy!?
  http://en.wikipedia.org/wiki/Watson_(computer)
- Helping President Obama win the election?
  http://www.rayidghani.com/
- Predicting the stock market with Twitter feed?
  http://arxiv.org/pdf/1010.3003&
- IEEE/ACM/AAAI fellow?

Seriously

- Do a great job in CSE 575!
- Read many many many … many papers
- Publish many many many … many papers
  - ICML: http://icml.cc/2017/
  - NIPS: http://nips.cc/Conferences/2017/
  - UAI: http://auai.org/uai2017/
  - IJCAI: http://ijcai17.org/
  - AAAI: http://www.aaai.org/Conferences/AAAI/aaai17.php
  - ACM KDD: http://www.kdd.org/kdd2017/
  - ICDM: http://icdm2017.bigke.org/
  - SDM: http://www.siam.org/meetings/sdm17/

  - Journal of Machine Learning Research: http://jmlr.org/
  - IEEE Transactions on Knowledge and Data Engineering: http://www.computer.org/portal/web/tkde

Who Wants Machine Learning People?

- IT
  - Outlier/fraud detection
  - Web image search
  - Recommendation
  - Information filtering
  - Community detection
  - Ad placement
  - Sentiment analysis
  - …
- Companies
  - Facebook, Google, LinkedIn, Twitter, Microsoft, IBM, AT&T, Apple, Amazon, Siemens, Foursquare, Yelp, Walmart Lab, NEC, Generic Electric, Baidu, Samsung, …

Who Wants Machine Learning People?

- Finance
  - Stock market prediction
  - Algorithmic trading
  - Return forecasting
  - …
- Companies
  - Goldman Sachs, Morgan Stanley, American Express, Citadel LLC, Barclays Capital, Rotella Capital Management, Citi Bank, Pequot Capital, Zestfinance, Federal Reserve Board, WorldQuant LLC, …

Who Wants Machine Learning People?

- Speech recognition, natural language processing
- Computer vision
- Healthcare
- Robot control
- Computational biology
- Sensor networks
- …

Basics on Probability

Coin Flips

- You flip a coin
  - Head with probability 0.5

- You flip 100 coins
  - How many heads would you expect

Coin Flips cont.

- You flip a coin
  - Head with probability *p*
  - Binary random variable
  - Bernoulli trial with success probability *p*
- You flip *k* coins
  - How many heads would you expect
  - Number of heads X: discrete random variable
  - Binomial distribution with parameters *k* and *p*

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g., the total number of heads X you get if you flip 100 coins
- X is a RV with arity *k* if it can take on exactly one value out of $\{x_1, \cdots, x_k\}$,
  - E.g., the possible values that X can take are 0, 1, 2,…, 100

Probability of Discrete RV

- Probability mass function (pmf): $\mathrm{P}\left(\mathrm{X}=x_i\right)$
- Easy facts about pmf
  - $\sum_i \mathrm{P}\left(\mathrm{X}=x_i\right)=1$
  - $\mathrm{P}\left(\mathrm{X}=x_i \cap \mathrm{X}=x_j\right)=0$ if $i \neq j$
  - $\mathrm{P}\left(\mathrm{X}=x_i \cup \mathrm{X}=x_j\right)=\mathrm{P}\left(\mathrm{X}=x_i\right)+\mathrm{P}\left(\mathrm{X}=x_j\right)$ if $i \neq j$
  - $\mathrm{P}(\mathrm{X}=x_1 \cup \mathrm{X}=x_2 \cup \cdots \cup \mathrm{X}=x_k)=1$

Common Distributions

- Uniform $X \sim U[1,\ldots,N]$
  - X takes values 1, 2, …, *N*
  - $P(X = i) = 1/N$
  - E.g., picking balls of different colors from a box
- Binomial $X \sim Bin(n, p)$
  - X takes values 0, 1, …, *n*
  - $P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$
  - E.g., coin flips

Coin Flips of Two Persons

- Your friend and you both flip coins
  - Head with probability 0.5
  - You flip 50 times; your friend flip 100 times
  - How many heads will both of you get

Joint Distribution

- Given two discrete RVs X and Y, their **joint distribution** is the distribution of X and Y together
  - E.g., P(You get 21 heads AND you friend get 70 heads)
- $$\sum_x \sum_y \mathrm{P}\left(\mathrm{X} = x \cap \mathrm{Y} = y\right) = 1$$
  - E.g.,

$$\sum_{i=0}^{50} \sum_{j=0}^{100} \mathrm{P}\left(\text{You get } i \text{ heads AND your friend get } j \text{ heads}\right) = 1$$

Conditional Probability

- $P\left(X=x \middle| Y=y\right)$ is the probability of $X=x$, given the occurrence of $Y=y$
  - E.g., you get 0 heads, given that your friend gets 61 heads
- $P\left(X=x \middle| Y=y\right)=\dfrac{P\left(X=x \cap Y=y\right)}{P\left(Y=y\right)}$

Law of Total Probability

- Given two discrete RVs X and Y, which take values in $\{x_1, \cdots, x_m\}$ and $\{y_1, \cdots, y_n\}$, we have

$$\mathrm{P}\left(\mathrm{X} = x_i\right) = \sum_j \mathrm{P}\left(\mathrm{X} = x_i \cap \mathrm{Y} = y_j\right)$$

$$= \sum_j \mathrm{P}\left(\mathrm{X} = x_i \middle| \mathrm{Y} = y_j\right)\mathrm{P}\left(\mathrm{Y} = y_j\right)$$

Marginalization

Marginal Probability

Joint Probability

$$
\mathrm{P}\left(\mathrm{X}=x_i\right)=\sum_j \mathrm{P}\left(\mathrm{X}=x_i \cap \mathrm{Y}=y_j\right)
$$

$$
=\sum_j \mathrm{P}\left(\mathrm{X}=x_i \middle| \mathrm{Y}=y_j\right)\mathrm{P}\left(\mathrm{Y}=y_j\right)
$$

Conditional Probability

Marginal Probability

Bayes Rule

- X and Y are discrete RVs…

$$P\left(X = x \middle| Y = y\right) = \frac{P\left(X = x \cap Y = y\right)}{P\left(Y = y\right)}$$

$$P\left(X = x_i \middle| Y = y_j\right) = \frac{P\left(Y = y_j \middle| X = x_i\right) P\left(X = x_i\right)}{\sum_k P\left(Y = y_j \middle| X = x_k\right) P\left(X = x_k\right)}$$

Independent RVs

- Intuition: X and Y are independent means that $\mathrm{X} = x$ **neither** makes it **more or less** probable that $\mathrm{Y} = y$
- Definition: X and Y are independent iff

$$\mathrm{P}\left(\mathrm{X} = x \cap \mathrm{Y} = y\right) = \mathrm{P}\left(\mathrm{X} = x\right)\mathrm{P}\left(\mathrm{Y} = y\right)$$

More on Independence

- $$P(X = x \cap Y = y) = P(X = x)P(Y = y)$$

$$P(X = x \mid Y = y) = P(X = x) \qquad P(Y = y \mid X = x) = P(Y = y)$$

- E.g., no matter how many heads you get, your friend will not be affected, and vice versa

Conditionally Independent RVs

- Intuition: X and Y are conditionally independent given Z means that once Z is **known**, the value of X does not add any **additional** information about Y
- Definition: X and Y are conditionally independent given Z iff

$$\mathrm{P}\left(\mathrm{X}=x \cap \mathrm{Y}=y \middle| \mathrm{Z}=z\right)=\mathrm{P}\left(\mathrm{X}=x \middle| \mathrm{Z}=z\right)\mathrm{P}\left(\mathrm{Y}=y \middle| \mathrm{Z}=z\right)$$

More on Conditional Independence

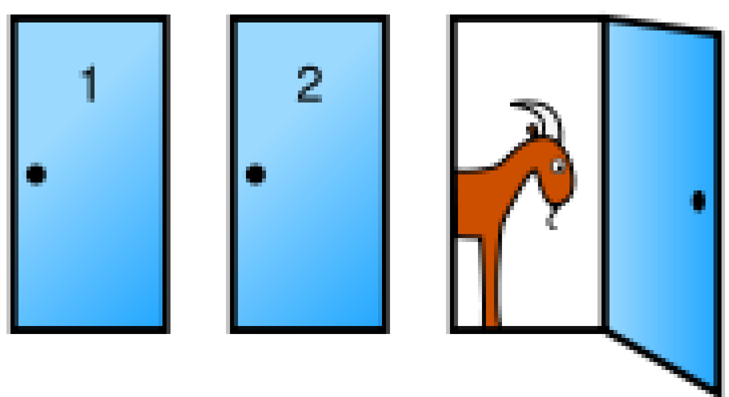
$$P\left(X = x \cap Y = y \middle| Z = z\right) = P\left(X = x \middle| Z = z\right) P\left(Y = y \middle| Z = z\right)$$

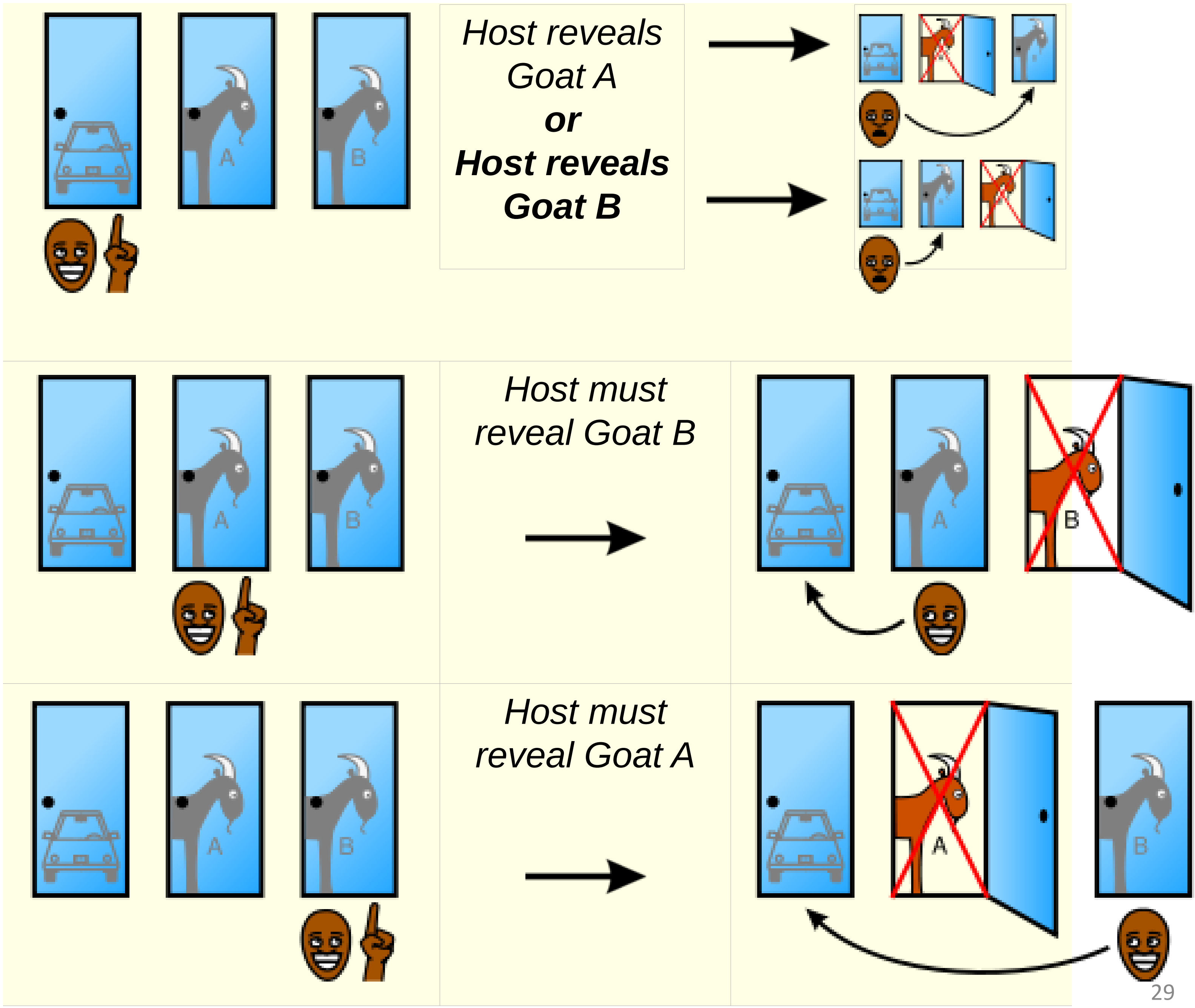
$$P\left(X = x \middle| Y = y, Z = z\right) = P\left(X = x \middle| Z = z\right)$$

$$P\left(Y = y \middle| X = x, Z = z\right) = P\left(Y = y \middle| Z = z\right)$$

Monty Hall Problem

- You're given the choice of three doors: Behind one door is a car; behind the others, goats.
- You pick a door, say No. 1
- The host, who knows what's behind the doors, opens another door, say No. 3, which has a goat.
- Do you want to pick door No. 2 instead?




Host reveals Goat A
or
Host reveals Goat B
Host must reveal Goat B
Host must reveal Goat A
A
B

Monty Hall Problem: Bayes Rule

- C_k: the car is behind door *k*, *k* = 1, 2, 3
- $P(C_k) = 1/3$
- H_{ij} : the host opens door *j* after you pick door *i*

- $$P(H_{ij} | C_k) = \begin{cases} 0 & i = j \\ 0 & j = k \\ 1/2 & i = k \\ 1 & i \neq k, j \neq k \end{cases}$$

Monty Hall Problem: Bayes Rule cont.

- WLOG, *i*=1 (your choice), *j*=3 (the host's choice)
- $P\left(C_1 \middle| H_{13}\right) = \dfrac{P\left(H_{13} \middle| C_1\right) P\left(C_1\right)}{P\left(H_{13}\right)}$
- $P\left(H_{13} \middle| C_1\right) P\left(C_1\right) = \dfrac{1}{2} \cdot \dfrac{1}{3} = \dfrac{1}{6}$

Monty Hall Problem: Bayes Rule cont.

- $$P(H_{13}) = P(H_{13}, C_1) + P(H_{13}, C_2) + P(H_{13}, C_3)$$
$$= P(H_{13} \mid C_1) P(C_1) + P(H_{13} \mid C_2) P(C_2)$$
$$= \frac{1}{6} + 1 \cdot \frac{1}{3}$$
$$= \frac{1}{2}$$
- $$P(C_1 \mid H_{13}) = \frac{1/6}{1/2} = \frac{1}{3}$$

Monty Hall Problem: Bayes Rule cont.

- $$P\left(C_1 \middle| H_{13}\right) = \frac{1/6}{1/2} = \frac{1}{3}$$
- $$P\left(C_2 \middle| H_{13}\right) = 1 - \frac{1}{3} = \frac{2}{3} > P\left(C_1 \middle| H_{13}\right)$$
- *You should switch!*

Continuous Random Variables

- What if X is continuous?
- Probability density function (pdf) instead of probability mass function (pmf)
- A pdf is any function $f\left(x\right)$ that describes the probability density in terms of the input variable *x*.

PDF

- Properties of pdf
  - $f(x) \geq 0, \forall x$
  - $\int_{-\infty}^{+\infty} f(x) = 1$
  - $f(x) \leq 1$???
- Actual probability can be obtained by taking the integral of pdf
  - E.g., the probability of X being between 0 and 1 is

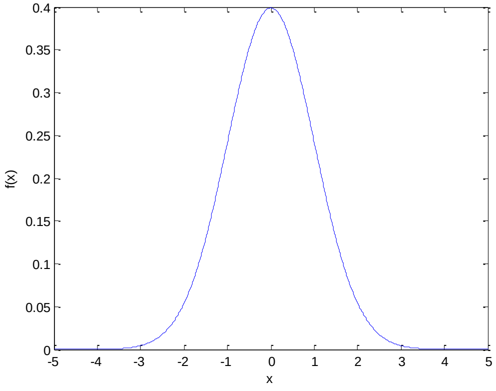
$$\mathrm{P}(0 \leq X \leq 1) = \int_0^1 f(x)dx$$

Cumulative Distribution Function

- $F_{\mathrm{X}}\left(v\right) = \mathrm{P}\left(\mathrm{X} \leq v\right)$
- Discrete RVs
  - $F_{\mathrm{X}}\left(v\right) = \sum_{v_i} \mathrm{P}\left(\mathrm{X} = v_i\right)$
- Continuous RVs
  - $F_{\mathrm{X}}\left(v\right) = \int_{-\infty}^{v} f\left(x\right) dx$
  - $\frac{d}{dx} F_{\mathrm{X}}\left(x\right) = f\left(x\right)$

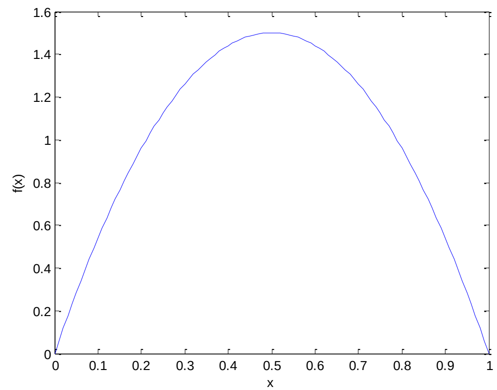
Common Distributions

- Normal $X \sim N(\mu, \sigma^2)$
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}, \quad x \in \Re$
  - E.g., the height of the entire population



Common Distributions cont.

- Beta $X \sim Beta(\alpha, \beta)$
  - $f\left(x;\alpha,\beta\right)=\frac{1}{B\left(\alpha,\beta\right)}x^{\alpha-1}\left(1-x\right)^{\beta-1},\ \ x\in\left[0,1\right]$
  - $\alpha=\beta=1$: uniform distribution between 0 and 1
  - E.g., the conjugate prior for the parameter *p* in Binomial distribution



Joint Distribution

- Given two continuous RVs X and Y, the **joint pdf** can be written as $f_{X,Y}(x, y)$

- $\int_x \int_y f_{X,Y}(x, y) dxdy = 1$

Multivariate Normal

- Generalization to higher dimensions of the one-dimensional normal

Covariance Matrix

- $$f_{\vec{X}}\left(x_1,\ldots,x_d\right)=\frac{1}{\left(2\pi\right)^{d/2}\left|\Sigma\right|^{1/2}}\cdot\exp\left\{-\frac{1}{2}\left(\vec{x}-\mu\right)^T\Sigma^{-1}\left(\vec{x}-\mu\right)\right\}$$

Mean

Moments

- Mean (Expectation): $\mu = E(\mathrm{X})$
  - Discrete RVs: $E(\mathrm{X}) = \sum_{v_i} v_i \mathrm{P}(\mathrm{X} = v_i)$
  - Continuous RVs: $E(\mathrm{X}) = \int_{-\infty}^{+\infty} x f(x)\, dx$
- Variance: $V(\mathrm{X}) = E(\mathrm{X} - \mu)^2$
  - Discrete RVs: $V(\mathrm{X}) = \sum_{v_i} (v_i - \mu)^2 \mathrm{P}(\mathrm{X} = v_i)$
  - Continuous RVs: $V(\mathrm{X}) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)\, dx$

Properties of Moments

- Mean
  - $E(\mathrm{X}+\mathrm{Y}) = E(\mathrm{X}) + E(\mathrm{Y})$
  - $E(a\mathrm{X}) = aE(\mathrm{X})$
  - If X and Y are independent, $E(\mathrm{XY}) = E(\mathrm{X}) \cdot E(\mathrm{Y})$
- Variance
  - $V(a\mathrm{X}+b) = a^2 V(\mathrm{X})$
  - If X and Y are independent, $V(\mathrm{X}+\mathrm{Y}) = V(\mathrm{X}) + V(\mathrm{Y})$

Moments of Common Distributions

- Uniform $X \sim U[1,\ldots,N]$
  - Mean $(1+N)/2$; variance $\left(N^2-1\right)/12$
- Binomial $X \sim Bin(n,p)$
  - Mean np; variance np^2
- Normal $X \sim N(\mu,\sigma^2)$
  - Mean μ; variance σ^2
- Beta $X \sim Beta(\alpha,\beta)$
  - Mean $\alpha/(\alpha+\beta)$; variance $\dfrac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Probability of Events

- X denotes an event that could possibly happen
  - E.g., X="you will fail in this course"
- P(X) denotes the **likelihood** that X happens, or X=true
  - E.g., what's the probability that you will fail in this course?
- Ω denotes the entire event set
  - $\Omega = \left\{ \mathrm{X}, \bar{\mathrm{X}} \right\}$

The Axioms of Probabilities

- 0 <= P(X) <= 1
- $\mathrm{P}(\Omega) = 1$
- $\mathrm{P}(\mathrm{X}_1 \cup \mathrm{X}_2 \cup \cdots) = \sum_i \mathrm{P}(\mathrm{X}_i)$, where X_i are disjoint events
- Useful rules
  - $\mathrm{P}(\mathrm{X}_1 \cup \mathrm{X}_2) = \mathrm{P}(\mathrm{X}_1) + \mathrm{P}(\mathrm{X}_2) - \mathrm{P}(\mathrm{X}_1 \cap \mathrm{X}_2)$
  - $\mathrm{P}(\bar{\mathrm{X}}) = 1 - \mathrm{P}(\mathrm{X})$

Interpreting the Axioms