Credits

- Slides adapted from Lee Cooper and Joydeep Ghosh
  - http://joyceho.github.io/cs534_s17/slide/2-prob-review.pdf
- UC Berkeley CS188 Intro to AI

- Also see
  - Appendix of the textbook.
  - http://cs229.stanford.edu/section/cs229-prob.pdf
  - http://cs229.stanford.edu/section/cs229-linalg.pdf

Applications

- Association
- Supervised Learning
    - Classification
    - Regression
- Unsupervised Learning
- Reinforcement Learning

Learning Associations

- Basket analysis:

  $P(Y \mid X)$ probability that somebody who buys X also buys Y where X and Y are products/services.

  Example: P (chips | beer) = 0.7

Regression

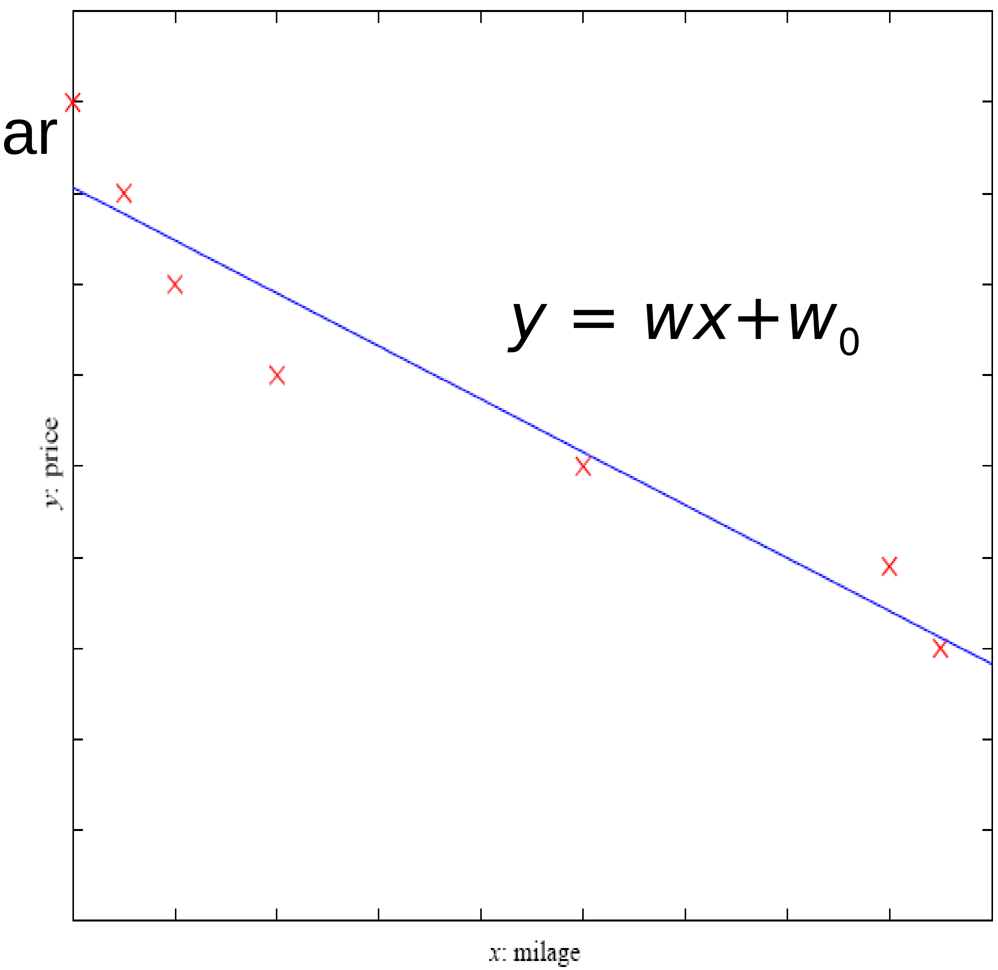
- Example: Price of a used car
- x : car attributes

  y : price

  $$y = g(x \mid \theta)$$

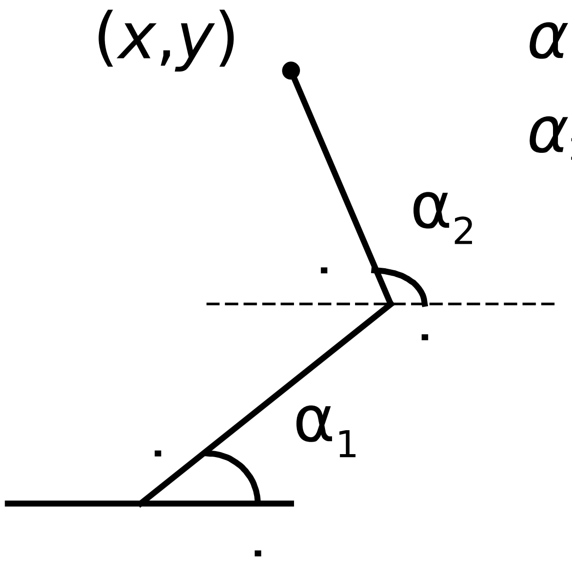
  g ( ) model,

θ parameters

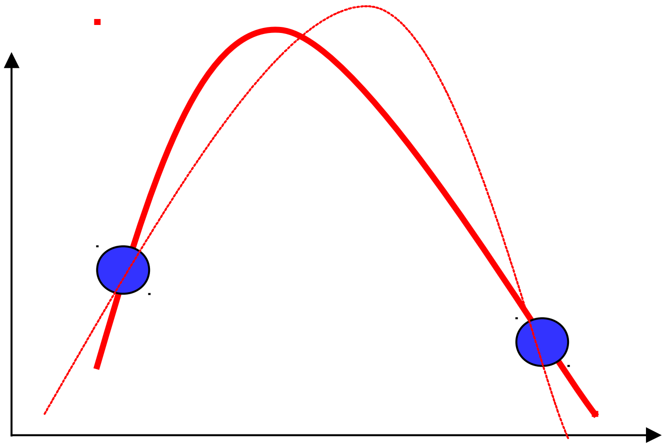


Regression Applications

- Navigating a car: Angle of the steering wheel
- Kinematics of a robot arm



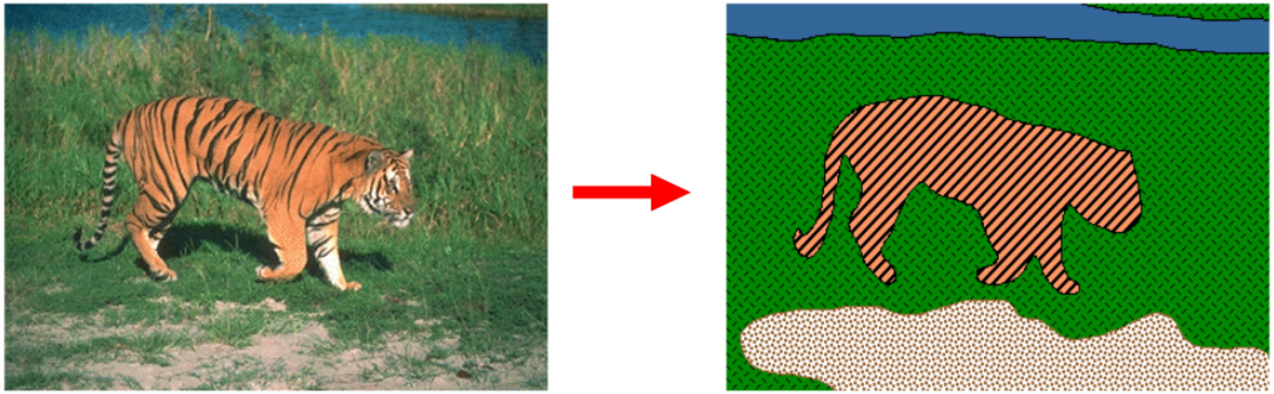
$$\alpha_1 = g_1(x,y)$$
$$\alpha_2 = g_2(x,y)$$



Unsupervised Learning

- Learning “what normally happens”
- No output
- Clustering: Grouping similar instances

Unsupervised Learning



- Biology : Discovering gene clusters with similar expression patterns, grouping homologous DNA sequences, etc.
- Marketing: Grouping customers with similar traits for segmenting the market, product positioning etc.
- Vision : Image segmentation, feature learning for recognition,...
- Social network analysis (discovering user communities with similar interests)
- Crime analysis (identification of “hot spots”)

Reinforcement Learning

- Learning a policy: A sequence of outputs
- No supervised output but delayed reward
- Credit assignment problem
- Game playing
- Robot in a maze
- Multiple agents, partial observability, ...

Resources: Datasets

- UCI Repository: http://www.ics.uci.edu/~mlearn/MLRepository.html
- UCI KDD Archive: http://kdd.ics.uci.edu/summary.data.application.html
- Statlib: http://lib.stat.cmu.edu/
- Delve: http://www.cs.utoronto.ca/~delve/

Resources: Journals

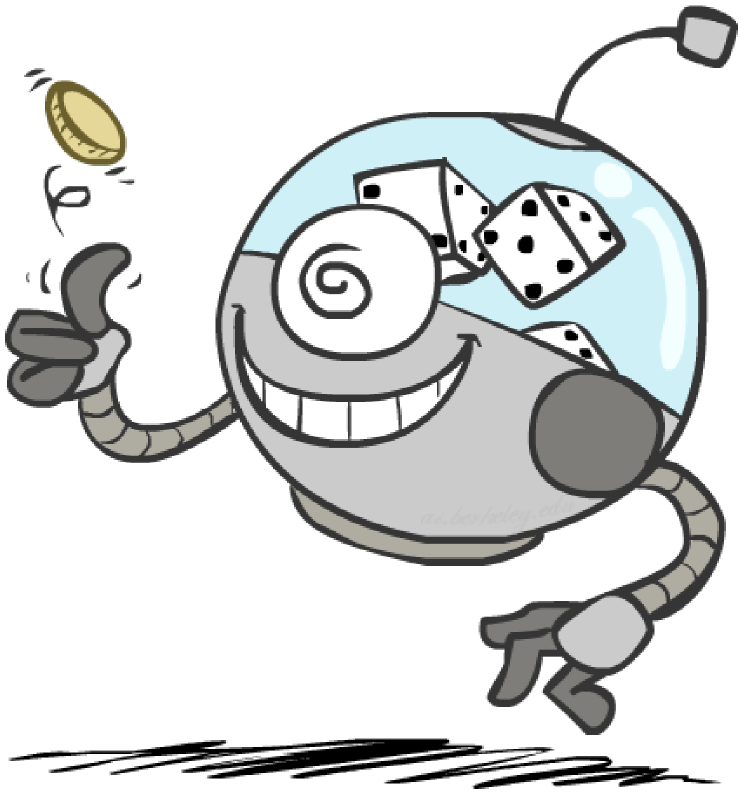
- Journal of Machine Learning Research www.jmlr.org
- Machine Learning
- Neural Computation
- Neural Networks
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association
- ...

Resources: Conferences

- International Conference on Machine Learning (ICML)
- ICML05: http://icml.ais.fraunhofer.de/
- European Conference on Machine Learning (ECML)
- ECML05: http://ecmlpkdd05.liacc.up.pt/
- Neural Information Processing Systems (NIPS)
- NIPS05: http://nips.cc/
- Uncertainty in Artificial Intelligence (UAI)
- UAI05: http://www.cs.toronto.edu/uai2005/
- Computational Learning Theory (COLT)
- COLT05: http://learningtheory.org/colt2005/
- International Joint Conference on Artificial Intelligence (IJCAI)
- IJCAI05: http://ijcai05.csd.abdn.ac.uk/
- International Conference on Neural Networks (Europe)
- ICANN05: http://www.ibspan.waw.pl/ICANN-2005/
- ...

Random Variables

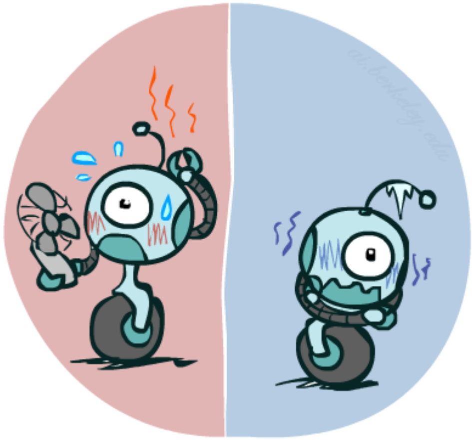
- A random variable is some aspect of the world about which w
  (may) have uncertainty
  - R = Is it raining?
  - T = Is it hot or cold?
  - D = How long will it take to drive to work?
  - L = Where is the ghost?
- We denote random variables with capital letters
- Like variables in a CSP, random variables have domains
  - R in {true, false}   (often write as {+r, -r})
  - T in {hot, cold}
  - D in [0, ∞)
  - L in possible locations, maybe {(0,0), (0,1), …}



Probability Distributions

- Associate a probability with each value

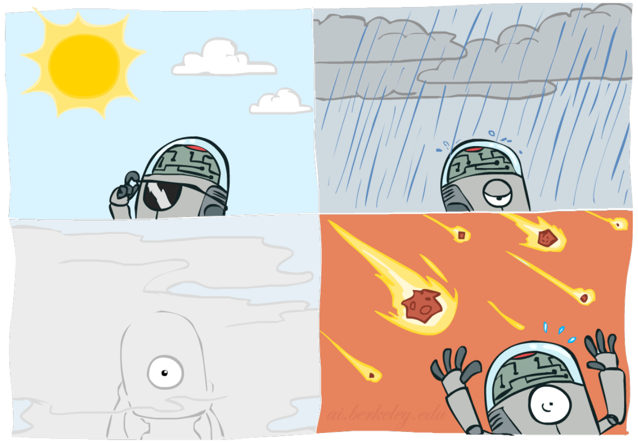
  - Temperature:



$P(T)$

T	P
hot	0.5
cold	0.5

  - Weather:



$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Probability Distributions

- Unobserved random variables have distributions

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

Shorthand notation:

$$P(hot) = P(T = hot),$$
$$P(cold) = P(T = cold),$$
$$P(rain) = P(W = rain),$$
$$\ldots$$

OK *if* all domain entries are unique

- A distribution is a TABLE of probabilities of values

- A probability (lower case value) is a single number

$$P(W = rain) = 0.1$$

- Must have: $\forall x \;\; P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Joint Distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$
  specifies a real number for each assignment (or *outcome*):

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

$$P(x_1, x_2, \ldots x_n) \geq 0$$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

Size of distribution if n variables with domain sizes d?

  - For all but the smallest distributions, impractical to write out!

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Prior probability

Prior or unconditional probabilities of propositions

e.g., $P(Cavity = true) = 0.1$ and $P(Weather = sunny) = 0.72$

correspond to belief prior to arrival of any (new) evidence

Probability distribution gives values for all possible assignments:

$\mathbf{P}(Weather) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$ (normalized, i.e., sums to 1)

Joint probability distribution for a set of r.v.s gives the probability of every atomic event on those r.v.s (i.e., every sample point)

$\mathbf{P}(Weather, Cavity) =$ a 4×2 matrix of values:

$Weather =$	$sunny$	$rain$	$cloudy$	$snow$
$Cavity = true$	0.144	0.02	0.016	0.02
$Cavity = false$	0.576	0.08	0.064	0.08

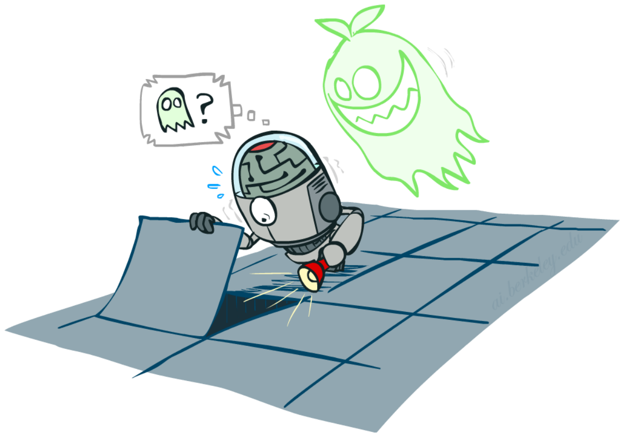
Every question about a domain can be answered by the joint distribution because every event is a sum of sample points

Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables
- Probabilistic models:
  - (Random) variables with domains
  - Assignments are called *outcomes*
  - Joint distributions: say whether assignments (outcomes) are likely
  - *Normalized:* sum to 1.0
  - Ideally: only certain variables directly interact
- Constraint satisfaction problems:
  - Variables with domains
  - Constraints: state whether assignments are possible
  - Ideally: only certain variables directly interact

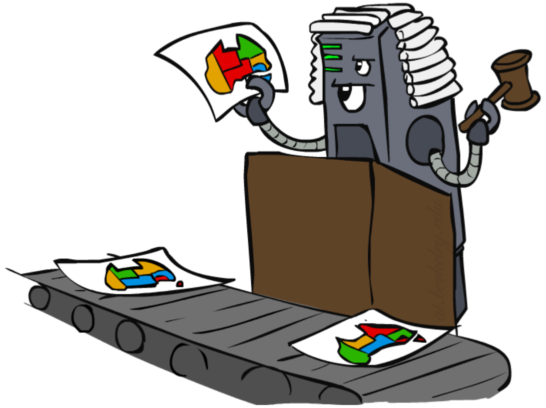
Distribution over T,W

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3



Constraint over T,W

T	W	P
hot	sun	T
hot	rain	F
cold	sun	F
cold	rain	T



Events

- An *event* is a set E of outcomes

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

- From a joint distribution, we can calculate the probability of any event
  - Probability that it's hot AND sunny?
  - Probability that it's hot?
  - Probability that it's hot OR sunny?
- Typically, the events we care about are *partial assignments*, like P(T=hot)

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Quiz: Events

- P(+x, +y) ?

- P(+x) ?

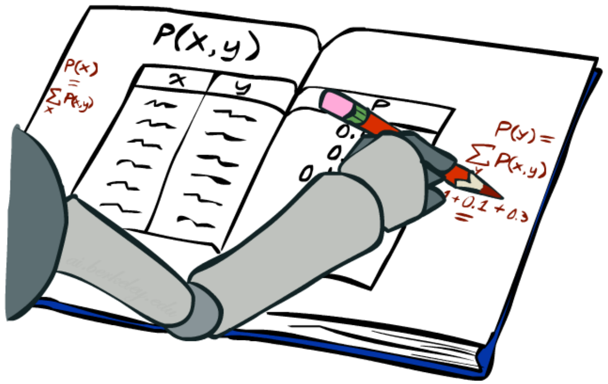
- P(-y OR +x) ?

$$P(X, Y)$$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding



$P(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(t) = \sum_s P(t, s)$$

$$P(s) = \sum_t P(t, s)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Quiz: Marginal Distributions

$P(X, Y)$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

$$P(x) = \sum_y P(x, y)$$

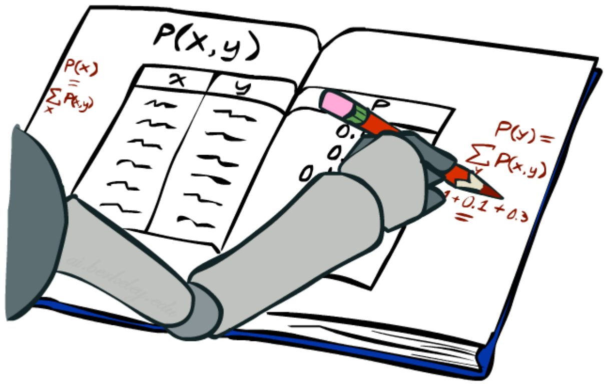
$$P(y) = \sum_x P(x, y)$$

$P(X)$

X	P
+x	
-x	

$P(Y)$

Y	P
+y	
-y	



Conditional probability

Conditional or posterior probabilities

e.g., $P(cavity|toothache) = 0.8$

i.e., **given that** $toothache$ **is all I know**

NOT "if $toothache$ then 80% chance of $cavity$"

(Notation for conditional distributions:

$\mathbf{P}(Cavity|Toothache) =$ 2-element vector of 2-element vectors)

If we know more, e.g., $cavity$ is also given, then we have

$P(cavity|toothache, cavity) = 1$

Note: the less specific belief **remains valid** after more evidence arrives, but is not always **useful**

New evidence may be irrelevant, allowing simplification, e.g.,

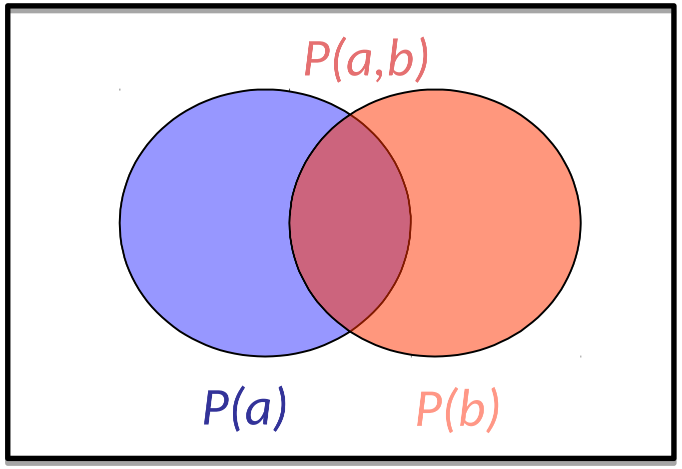
$P(cavity|toothache, 49ersWin) = P(cavity|toothache) = 0.8$

This kind of inference, sanctioned by domain knowledge, is crucial

Conditional Probabilities

- A simple relation between joint and conditional probabilities
  - In fact, this is taken as the *definition* of a conditional probability

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$$P(T,W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$P(W=s|T=c) = \frac{P(W=s,T=c)}{P(T=c)} = \frac{0.2}{0.5} = 0.4$$

$$= P(W=s,T=c) + P(W=r,T=c)$$

$$= 0.2 + 0.3 = 0.5$$

Quiz: Conditional Probabilities

$$P(X, Y)$$

X	Y	P
+x	+y	0.2
+x	-y	0.3
-x	+y	0.4
-x	-y	0.1

- P(+x | +y) ?
- P(-x | +y) ?
- P(-y | +x) ?

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Conditional Distributions

$$P(W|T)$$

$$P(W|T = hot)$$

W	P
sun	0.8
rain	0.2

$$P(W|T = cold)$$

W	P
sun	0.4
rain	0.6

Joint Distribution

$$P(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Inference by enumeration

Start with the joint distribution:

	$toothache$		$\neg\, toothache$	
	$catch$	$\neg\, catch$	$catch$	$\neg\, catch$
$cavity$	**.108**	**.012**	**.072**	**.008**
$\neg\, cavity$	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:
$P(\phi) = \Sigma_{\omega:\omega\models\phi} P(\omega)$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \Sigma_{\omega:\omega\models\phi} P(\omega)$$

$$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

Inference by enumeration

Start with the joint distribution:

	$toothache$		$\neg\, toothache$	
	$catch$	$\neg\, catch$	$catch$	$\neg\, catch$
$cavity$	**.108**	**.012**	**.072**	**.008**
$\neg\, cavity$	**.016**	**.064**	**.144**	**.576**

For any proposition ϕ, sum the atomic events where it is true:

$P(\phi) = \Sigma_{\omega:\omega\models\phi} P(\omega)$

$P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$

Inference by enumeration

Start with the joint distribution:

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ *cavity*	**.016**	**.064**	**.144**	**.576**

Can also compute conditional probabilities:

$$P(\neg cavity | toothache) = \frac{P(\neg cavity \wedge toothache)}{P(toothache)}$$

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Normalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

Denominator can be viewed as a normalization constant α

$$\begin{aligned}
&\mathbf{P}(Cavity|toothache) = \alpha\,\mathbf{P}(Cavity, toothache)\\
&\quad = \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]\\
&\quad = \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle]\\
&\quad = \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle
\end{aligned}$$

General idea: compute distribution on query variable
by fixing evidence variables and summing over hidden variables

Inference by enumeration, contd.

Let $\mathbf{X}$ be all the variables. Typically, we want
the posterior joint distribution of the query variables $\mathbf{Y}$
given specific values $\mathbf{e}$ for the evidence variables $\mathbf{E}$

Let the hidden variables be $\mathbf{H} = \mathbf{X} - \mathbf{Y} - \mathbf{E}$

Then the required summation of joint entries is done by summing out the hidden variables:

$$\mathbf{P}(\mathbf{Y}|\mathbf{E}=\mathbf{e}) = \alpha\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e}) = \alpha\Sigma_{\mathbf{h}}\mathbf{P}(\mathbf{Y},\mathbf{E}=\mathbf{e},\mathbf{H}=\mathbf{h})$$

The terms in the summation are joint entries because $\mathbf{Y}$, $\mathbf{E}$, and $\mathbf{H}$ together exhaust the set of random variables

Obvious problems:
1) Worst-case time complexity $O(d^n)$ where d is the largest arity
2) Space complexity $O(d^n)$ to store the joint distribution
3) How to find the numbers for $O(d^n)$ entries???

Inference by Enumeration

- P(W)?

- P(W | winter)?

- P(W | winter, hot)?

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

Random Variable Types

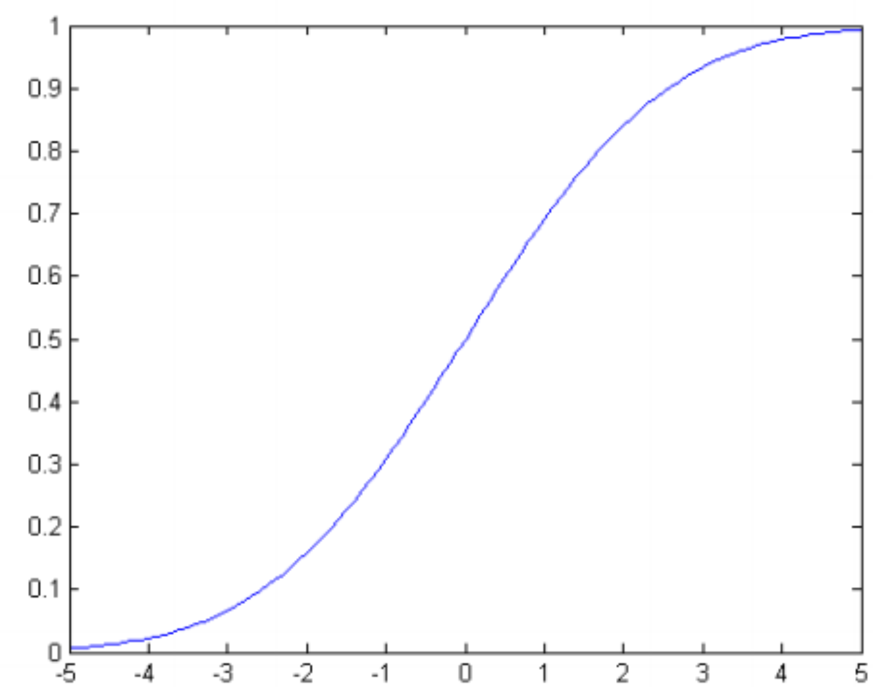
- Discrete random variable: X can take only a finite number of values
  - Example: Number of heads in a sequence of tosses
- Continuous random variable: X takes infinite number of possible values
  - Example: Amount of time for a radioactive particle to decay

Cumulative Distribution Function (CDF)

- CDF: function $F_X : \mathbb{R} \rightarrow [0, 1]$ that specifies a probability measure

$$F_X(x) \triangleq P(X \leq x)$$

- Used to calculate the probability of an event in $\mathcal{F}$
- Properties:
  - $0 \leq F_X(x) \leq 1$
  - $\lim_{x \rightarrow -\infty} F_X(x) = 0$
  - $\lim_{x \rightarrow \infty} F_X(x) = 1$
  - $x \leq y \implies F_X(x) \leq F_X(y)$

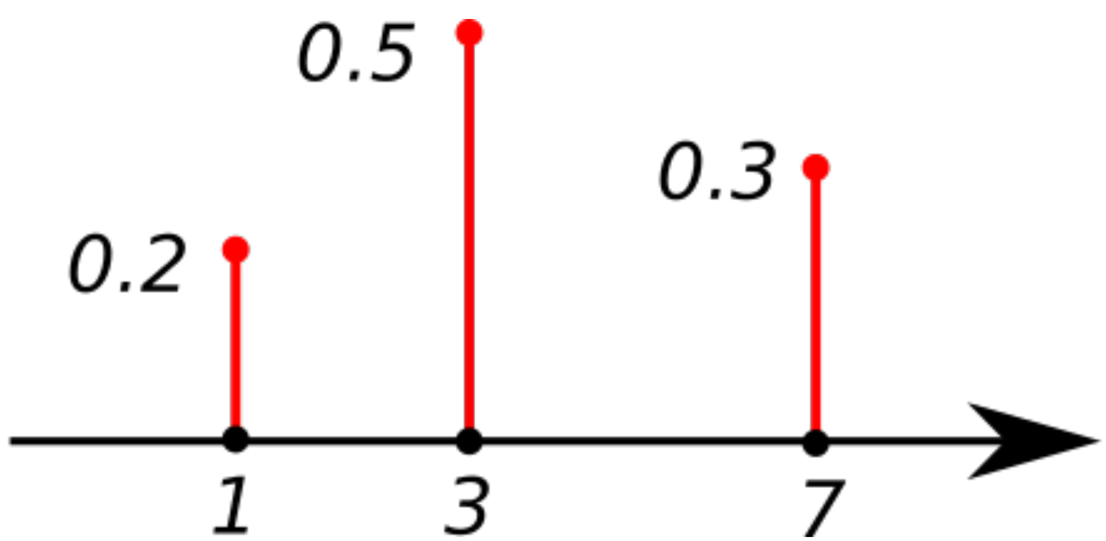


Probability Mass Function (PMF)

- Probability measure for discrete random variable
- PMF: function $p_X(x) : \Omega \to \mathbb{R}$ such that

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

- Properties:
  - $0 \leq p_X(x) \leq 1$
  - $\sum_{x \in \mathrm{Val}(X)} P_X(x) = 1$
  - $\sum_{x \in A} P_X(x) = P(X \in A)$



https://en.wikipedia.org/wiki/Probability_mass_function

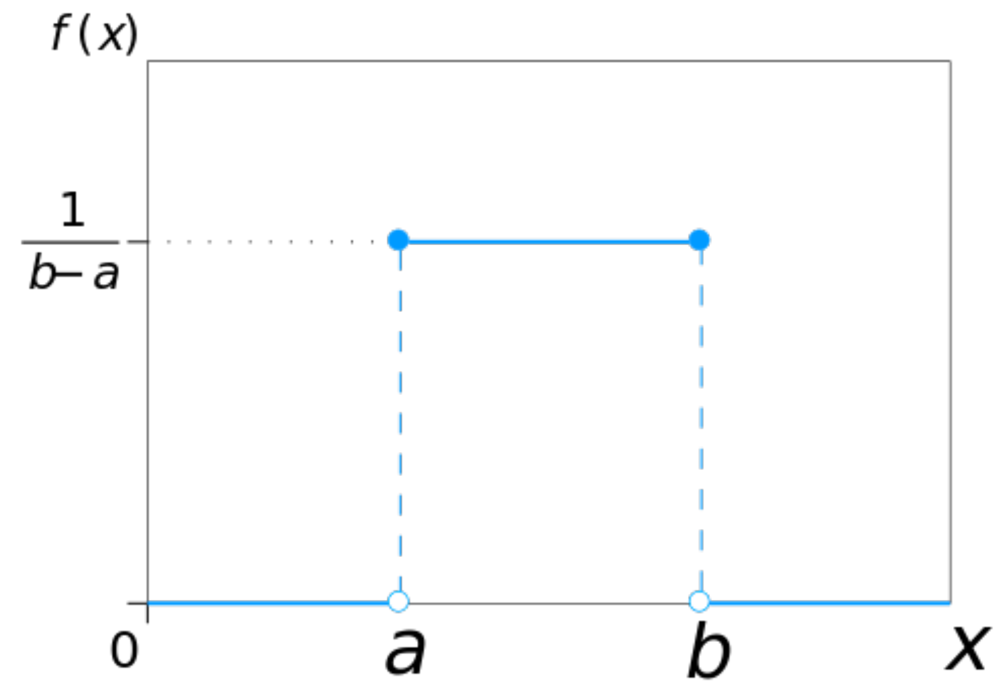
Probability Density Function (PDF)

- Probability measure for continuous random variable
- PDF is derivative of CDF

$$f_X(x) \triangleq \frac{dF_X(x)}{dx}$$

may not always exist if CDF is not differentiable

- Properties:
  - $f_X(x) \geq 0$
  - $\int_{-\infty}^{\infty} f_X(x) dx = 1$
  - $\int_{x \in A} f_X(x) dx = P(X \in A)$



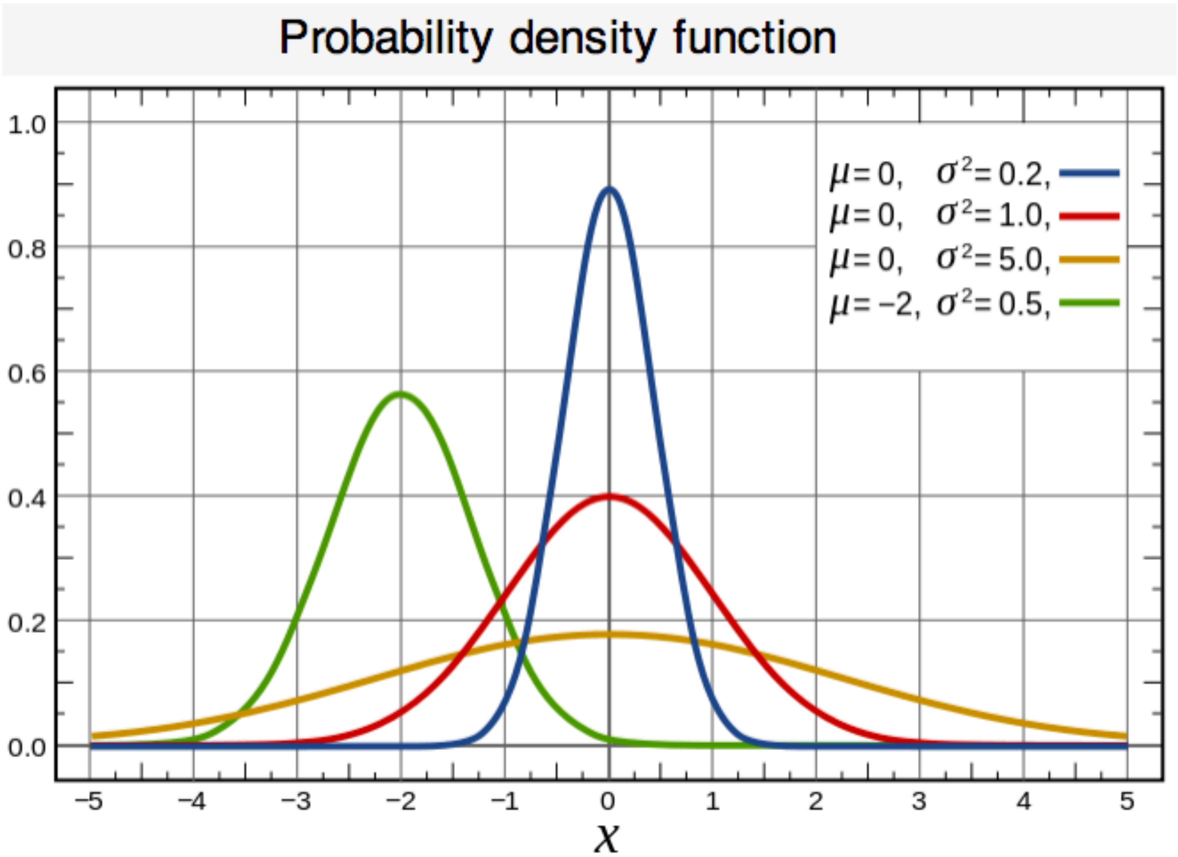
https://en.wikipedia.org/wiki/Uniform_distribution_(continuous)

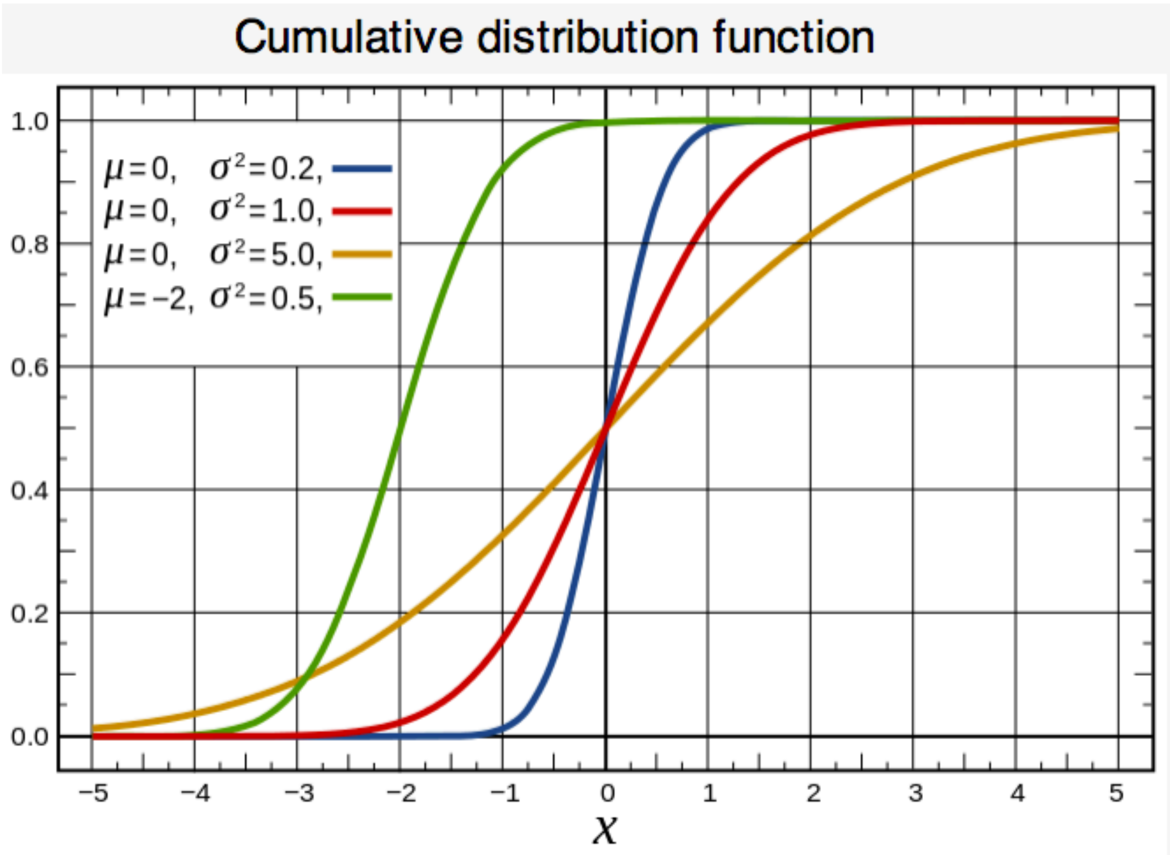
Example: Normal Distribution

mean

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

variance





https://en.wikipedia.org/wiki/Normal_distribution

PDFs vs. PMFs

	PDF	PMF
Values	Continuous valued RVs	Discrete-valued RVs
Representation	Function f(x)	Table
Probability	Calculated via integration	Calculated via summation
P(x = k)	0	Non-zero

Expectation: Mean and Variance

Expectation

- What is the expected value of a random variable?

- Expectation of g(X):

$$E[g(X)] \triangleq \sum_{x \in \mathrm{Val}(X)} g(x) p_X(X)$$

$$E[g(X)] \triangleq \int_{-\infty}^{\infty} g(x) f_X(X)$$

- "Weighted average" of values that g(x) with weights given by pdf or pmf

Expectation: Properties

- Constant

$$E[a] = a, \; a \in \mathbb{R}$$

- Scalar

$$E[af(X)] = aE[f(X)], \; a \in \mathbb{R}$$

- Linearity

$$E[f(X) + g(X)] = E[f(X)] + E[g(X)]$$

Expectation: Common Forms

- Mean: expectation of random variable

$$E[X], \text{ where } g(x) = x$$

- Variance: measure of how concentrated the distribution of the random variable is around its mean

$$\mathrm{Var}[X] \triangleq E[(X - E[X])^2]$$

Common Distributions

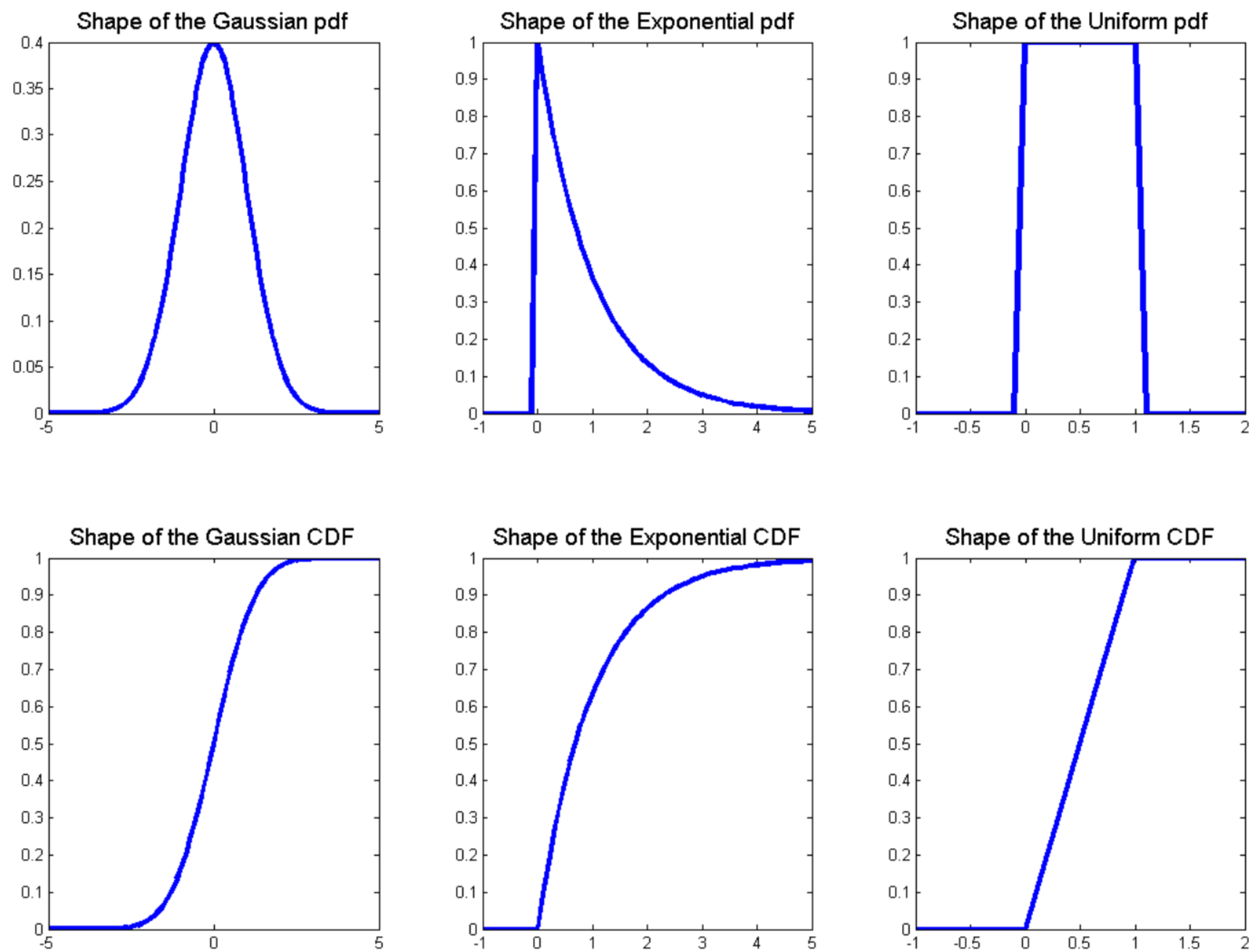
Discrete RV Distributions

- Bernoulli(p): coin flip with probability p of getting a heads (p = 1)
- Binomial(n,p): number of heads in n independent flips of a coin with probability p of a heads
- Geometric(p): number of flips of a coin until the first heads
- Poisson(λ): frequency of events or counts

Continuous RV Distributions

- Uniform(a, b): equal probability density between every value a and b on the real line
- Exponential(λ): decaying probability density over the nonnegative real numbers
- Normal(μ, σ^2): Gaussian distribution
  - Will be dealing with this 99% of the time
  - Interesting properties

Continuous RVs: PDF & CDF



http://cs229.stanford.edu/section/cs229-prob.pdf

Common RV Summary

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $0 \leq k \leq n$	np	npq
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$e^{-\lambda}\lambda^x/x!$ for $k = 1, 2, \ldots$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a}$ $\forall x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ $x \geq 0, \lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

http://cs229.stanford.edu/section/cs229-prob.pdf