Probability, Decision Theory, and Loss Functions

CMSC 678

UMBC

Some slides adapted from Hamed Pirsiavash

A Terminology Buffet

Classification

Regression

Clustering

the ***task****: what* kind *of problem are you solving?*

A Terminology Buffet

Classification	Fully-supervised
Regression	Semi-supervised
Clustering	Un-supervised
the ***task****: what* kind *of problem are you solving?*	*the* ***data****: amount of human input/number of labeled examples*

A Terminology Buffet

Classification	Fully-supervised	Probabilistic	Neural
Regression	Semi-supervised	Generative	Memory-based
Clustering	Un-supervised	Conditional	Exemplar
		Spectral	…
*the **task**: what* kind *of problem are you solving?*	*the **data**: amount of human input/number of labeled examples*	*the **approach**: how any data are being used*	

Outline

Review+Extension

Probability

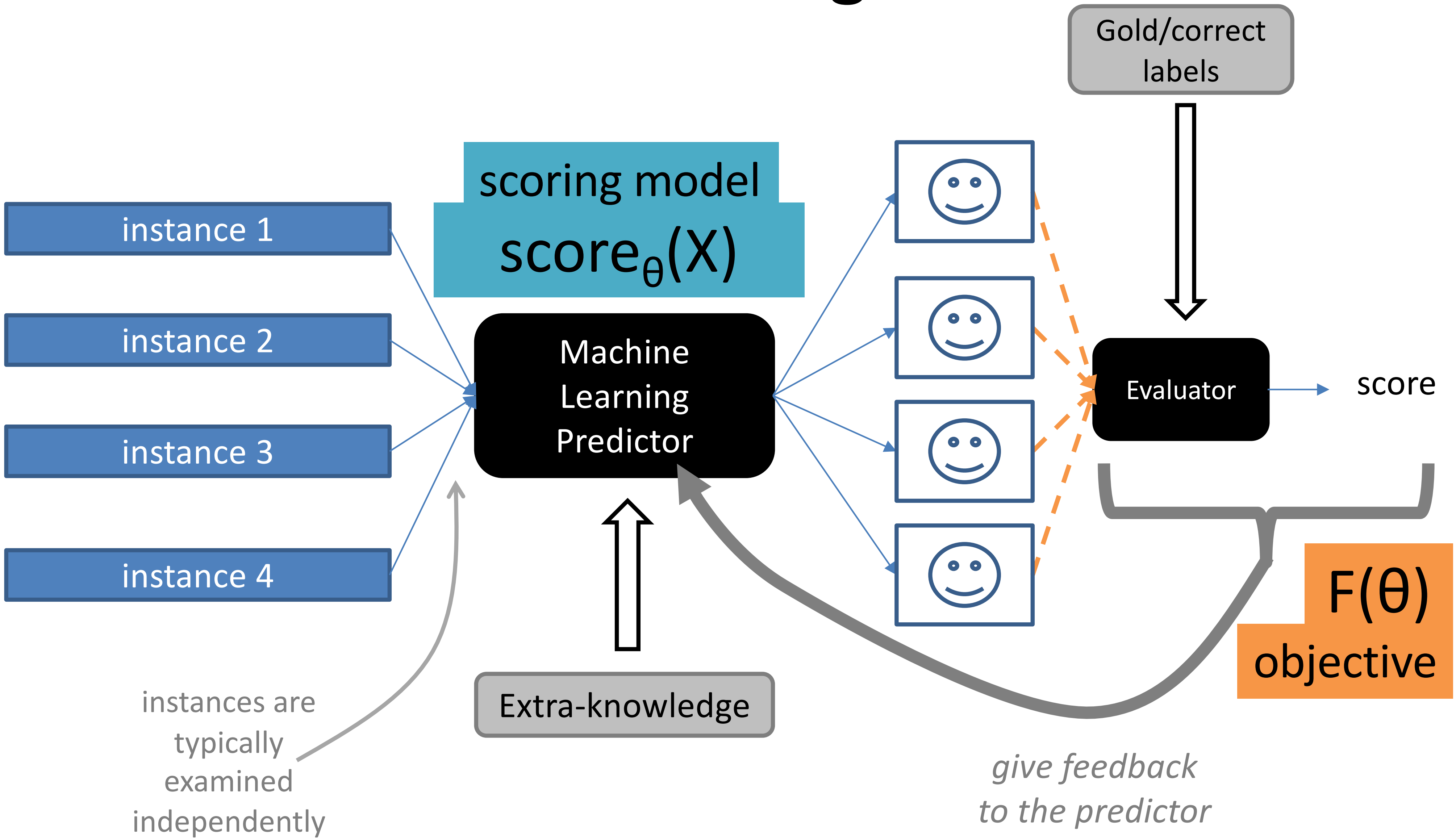
Decision Theory

Loss Functions

What does it mean to learn?

Generalization

Machine Learning Framework: Learning



Model, parameters and hyperparameters

Model: mathematical formulation of system (e.g., classifier)

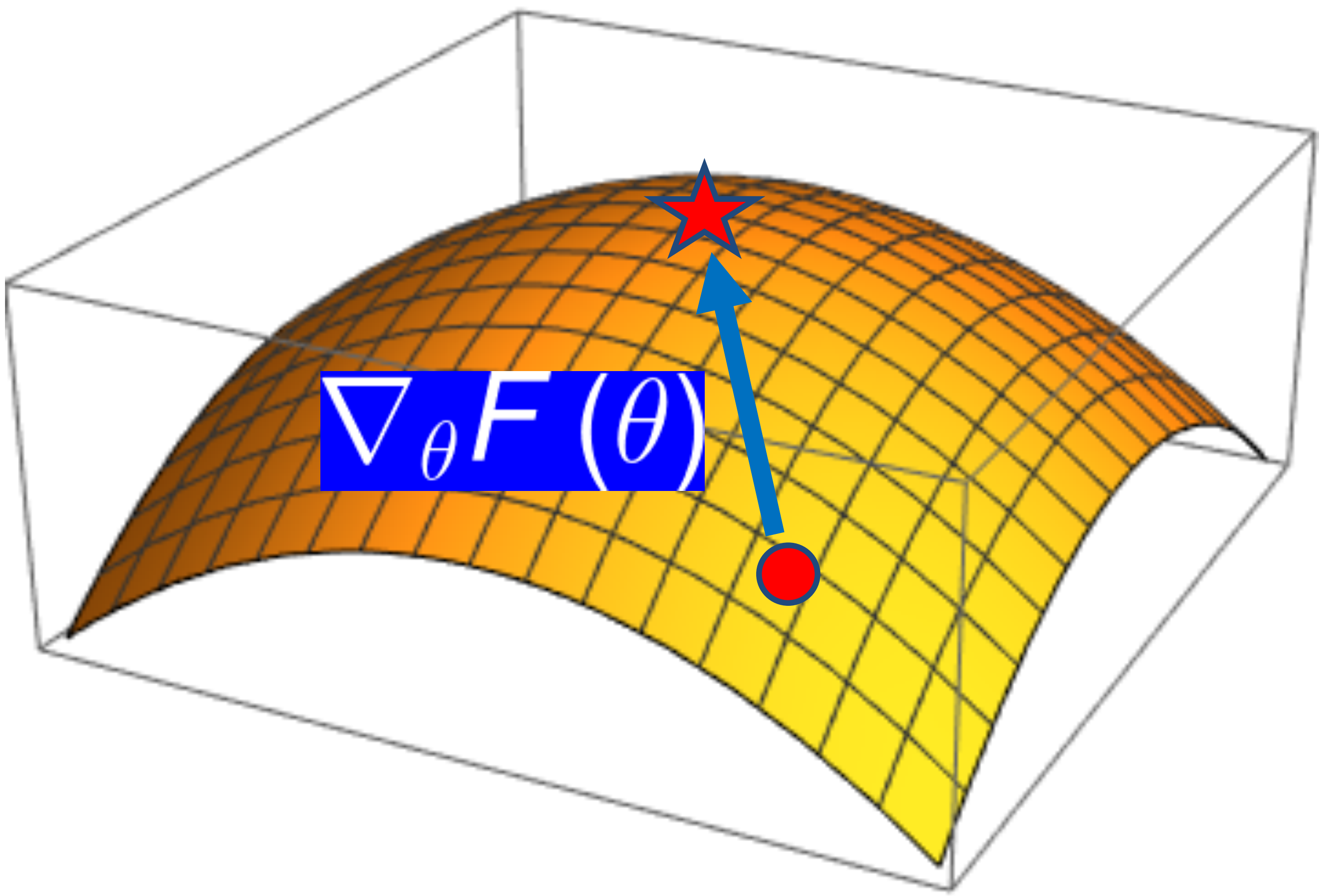
Parameters: primary "knobs" of the model that are set by a learning algorithm

Hyperparameter: secondary "knobs"

http://www.uiparade.com/wp-content/uploads/2012/01/ui-design-pure-css.jpg

Gradient Ascent

$$\arg\max_{\theta} \; F(\theta)$$



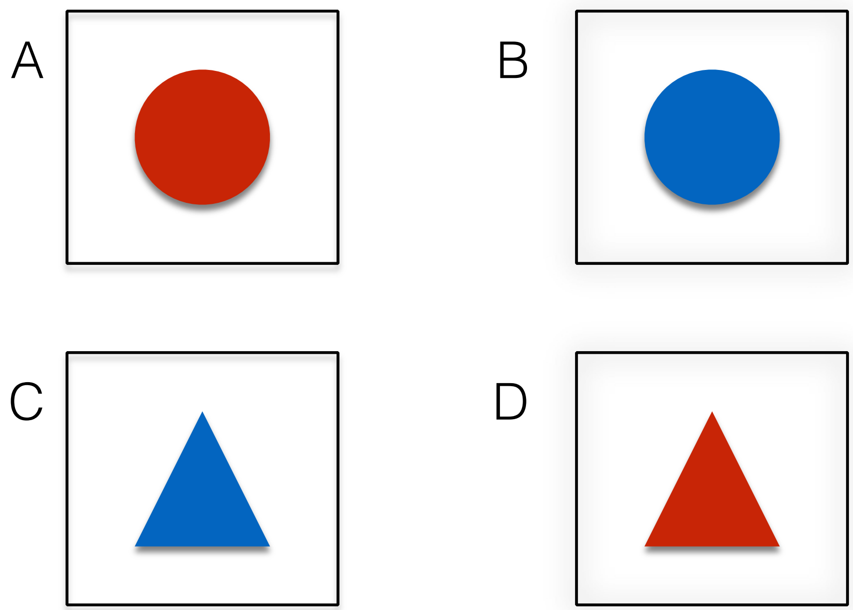
General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?

Courtesy Hamed Pirsiavash

General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?

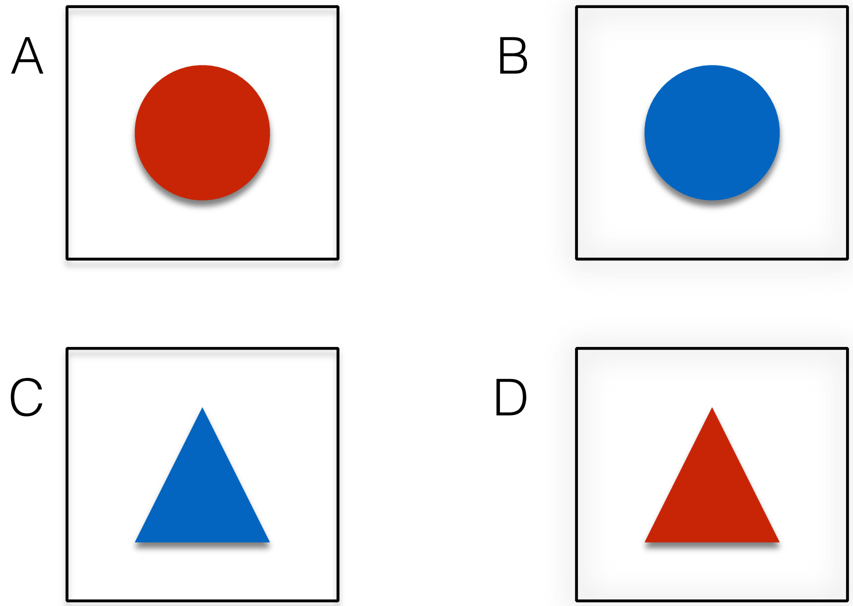


Partition these into two groups…

Courtesy Hamed Pirsiavash

General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?



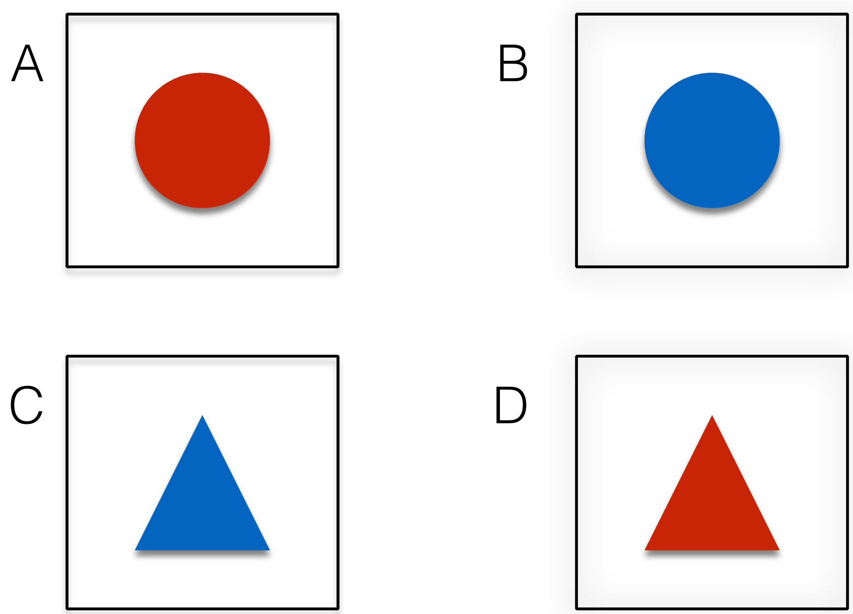
Partition these into two groups

Who selected red vs. blue?

Courtesy Hamed Pirsiavash

General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?



Partition these into two groups

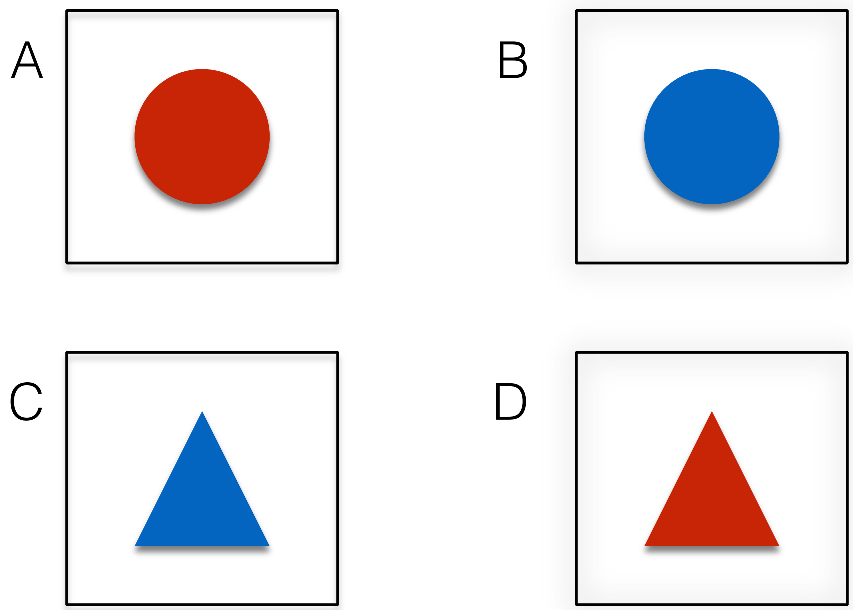
Who selected red vs. blue?

Who selected ● vs. ▲ ?

Courtesy Hamed Pirsiavash

General ML Consideration: Inductive Bias

What do we know *before* we see the data, and how does that influence our modeling decisions?



Partition these into two groups

Who selected red vs. blue?

Who selected ● vs. ▲ ?

Tip: Remember how your own biases/interpretation are influencing your approach

Courtesy Hamed Pirsiavash

Today’s Goals:

1. Remember Probability/Statistics
2. Understand Optimizing Empirical Risk

Outline

Review+Extension

Probability

Decision Theory

Loss Functions

Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable

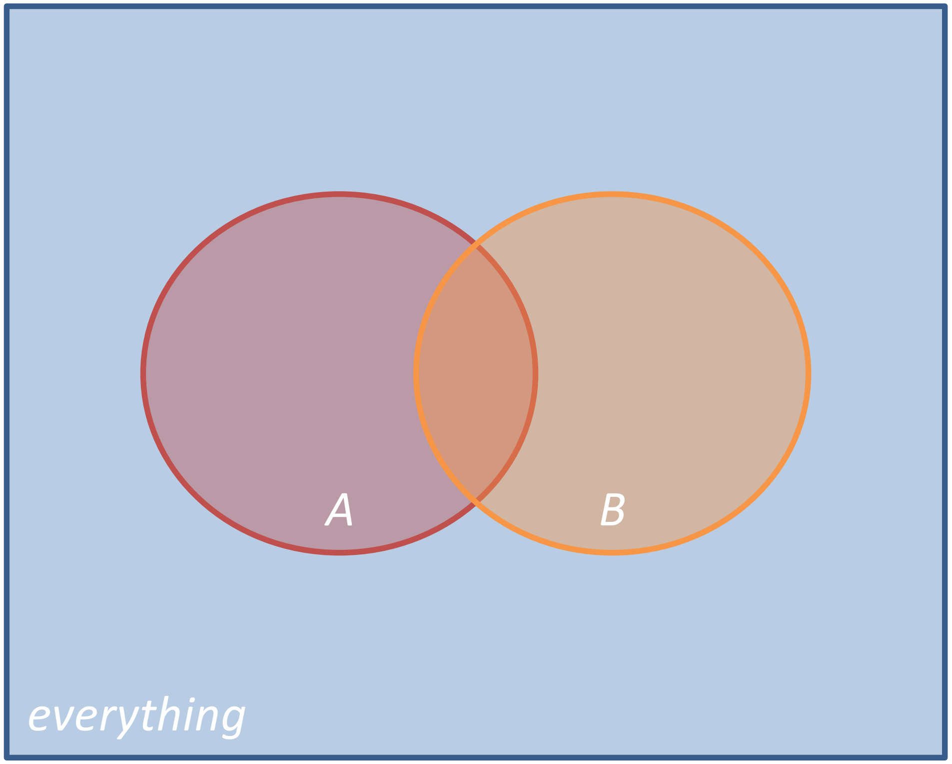
(Most) Probability Axioms

$p(\text{everything}) = 1$

$p(\phi) = 0$

$p(A) \leq p(B)$, when $A \subseteq B$

$p(A \cup B) = p(A) + p(B)$,
when $A \cap B = \phi$



$p(A \cup B) \neq p(A) + p(B)$

$p(A \cup B) = p(A) + p(B) - p(A \cap B)$

Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random "process"

Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”

Example #1: A (weighted) coin that can come up heads or tails

- X is a random variable denoting the possible outcomes
- X=HEADS or X=TAILS

Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random “process”

Example #1: A (weighted) coin that can come up heads or tails

- X is a random variable denoting the possible outcomes
- X=HEADS or X=TAILS

Example #2: Measuring the amount of snow that fell in the last storm

- Y is a random variable denoting the amount snow that fell, in inches
- Y=0, or Y=0.5, or Y=1.0495928591, or Y=10, or …

Probabilities and Random Variables

Random variables: variables that represent the possible outcomes of some random "process"

Example #1: A (weighted) coin that can come up heads or tails

X is a random variable denoting the possible outcomes

X=HEADS or X=TAILS

DISCRETE random variable

Example #2: Measuring the amount of snow that fell in the last storm

Y is a random variable denoting the amount snow that fell, in inches

Y=0, or Y=0.5, or Y=1.0495928591, or Y=10, or …

CONTINUOUS random variable

Random Variables

	If X is a…	
	Discrete random variable	Continuous random variable
The values k that X can take are	Discrete: finite or countably infinite (e.g., integers)	Continuous: uncountably infinite (e.g., real values)

Random Variables

	If X is a…	
	Discrete random variable	Continuous random variable
The values k that X can take are	Discrete: finite or countably infinite (e.g., integers)	Continuous: uncountably infinite (e.g., real values)
The function that gives the relative likelihood of a value p(X=k) is a	probability mass function (PMF)	probability density function (PDF)

Random Variables

	If X is a…	
	Discrete random variable	Continuous random variable
The values k that X can take are	Discrete: finite or countably infinite (e.g., integers)	Continuous: uncountably infinite (e.g., real values)
The function that gives the relative likelihood of a value p(X=k) is a	probability mass function (PMF)	probability density function (PDF)
The values that PMF/PDF can take are	$0 \leq p(X=k) \leq 1$	$p(X=k) \geq 0$

Random Variables

	If X is a…	
	Discrete random variable	Continuous random variable
The values k that X can take are	Discrete: finite or countably infinite (e.g., integers)	Continuous: uncountably infinite (e.g., real values)
The function that gives the relative likelihood of a value p(X=k) is a	probability mass function (PMF)	probability density function (PDF)
The values that PMF/PDF can take are	$0 \leq p(X=k) \leq 1$	$p(X=k) \geq 0$
We “add” with	Sums ($\sum$)	Integrals ($\int$)

Random Variables

	If X is a…	
	Discrete random variable	Continuous random variable
The values k that X can take are	Discrete: finite or countably infinite (e.g., integers)	Continuous: uncountably infinite (e.g., real values)
The function that gives the relative likelihood of a value p(X=k) is a	probability mass function (PMF)	probability density function (PDF)
The values that PMF/PDF can take are	$0 \leq p(X=k) \leq 1$	$p(X=k) \geq 0$
We “add” with	Sums ($\sum$)	Integrals ($\int$)
Our PMF/PDF satisfies p(everything)=1 by	$\sum_k p(X=k) = 1$	$\int p(x)dx = 1$

Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

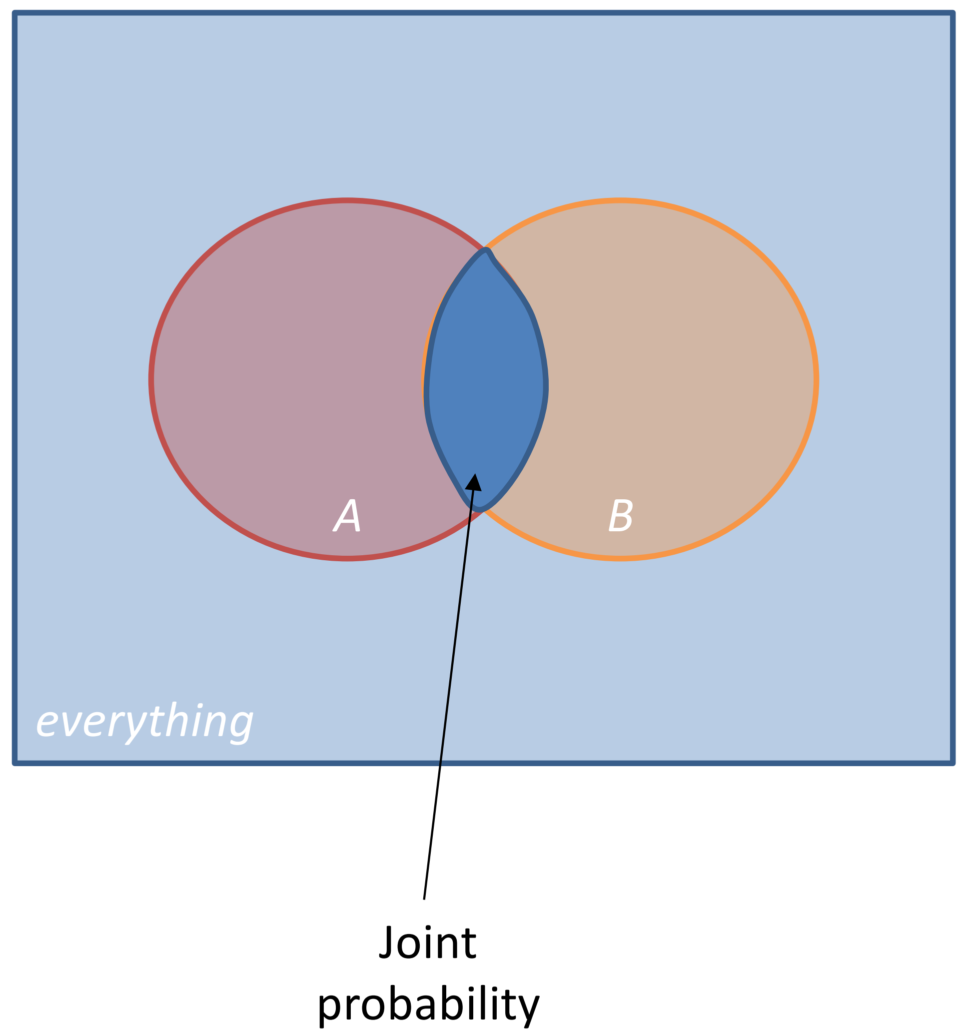
Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable

Joint Probability

Probability that multiple things "happen together"

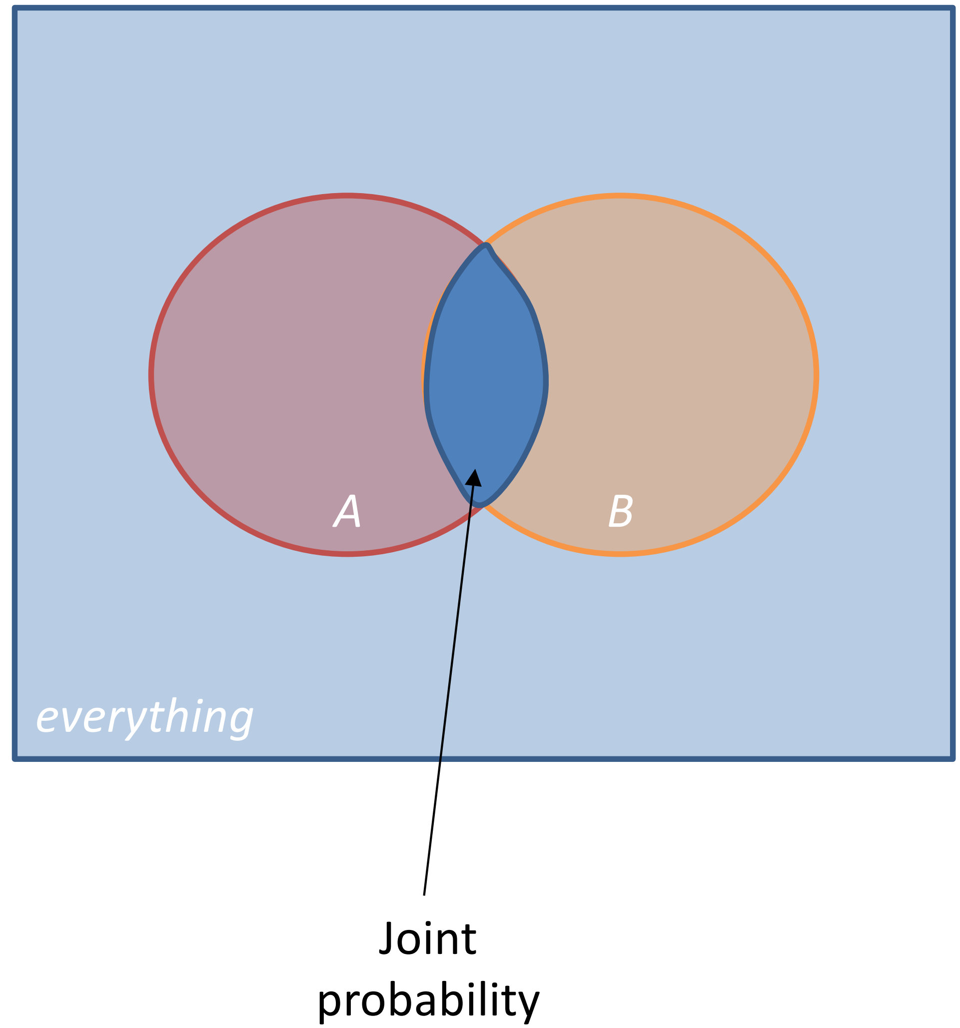


Joint Probability

Probability that multiple things "happen together"

$p(x,y)$, $p(x,y,z)$, $p(x,y,w,z)$

Symmetric: $p(x,y) = p(y,x)$



Joint Probability

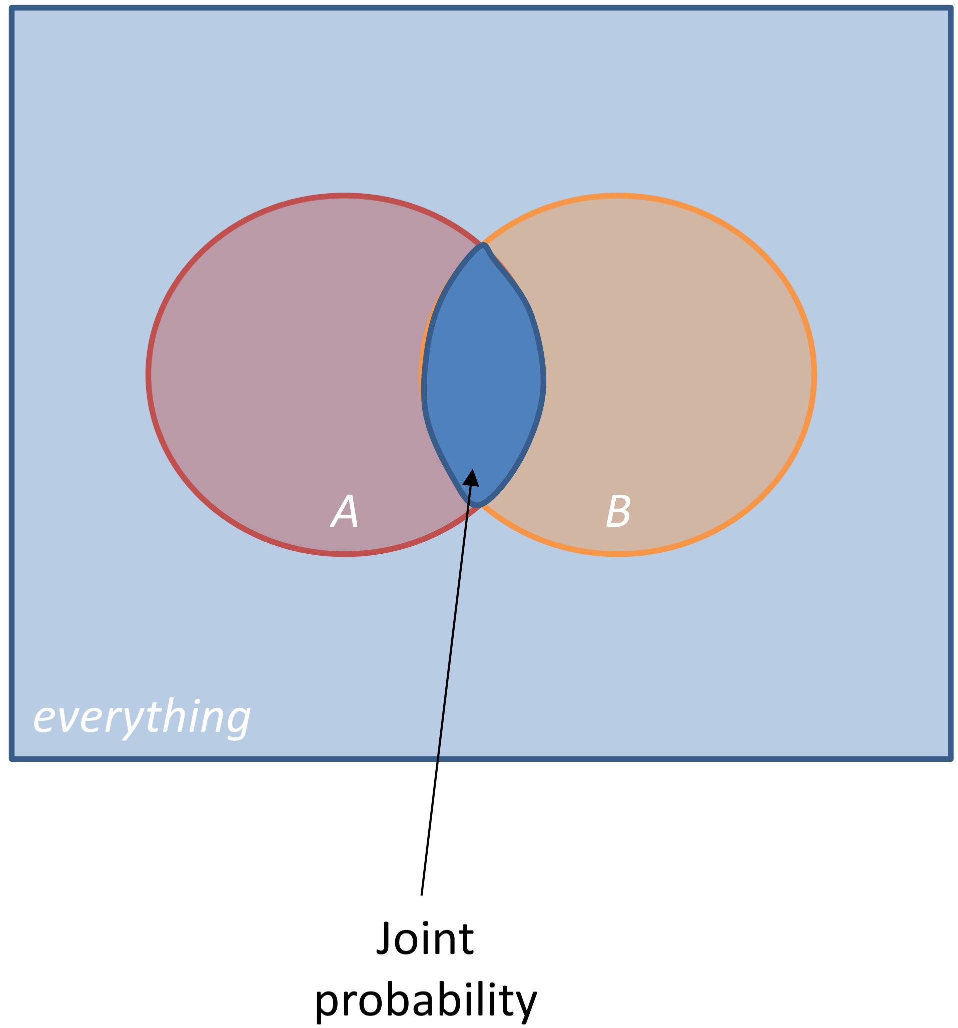
Probability that multiple things "happen together"

p(x,y), p(x,y,z), p(x,y,w,z)

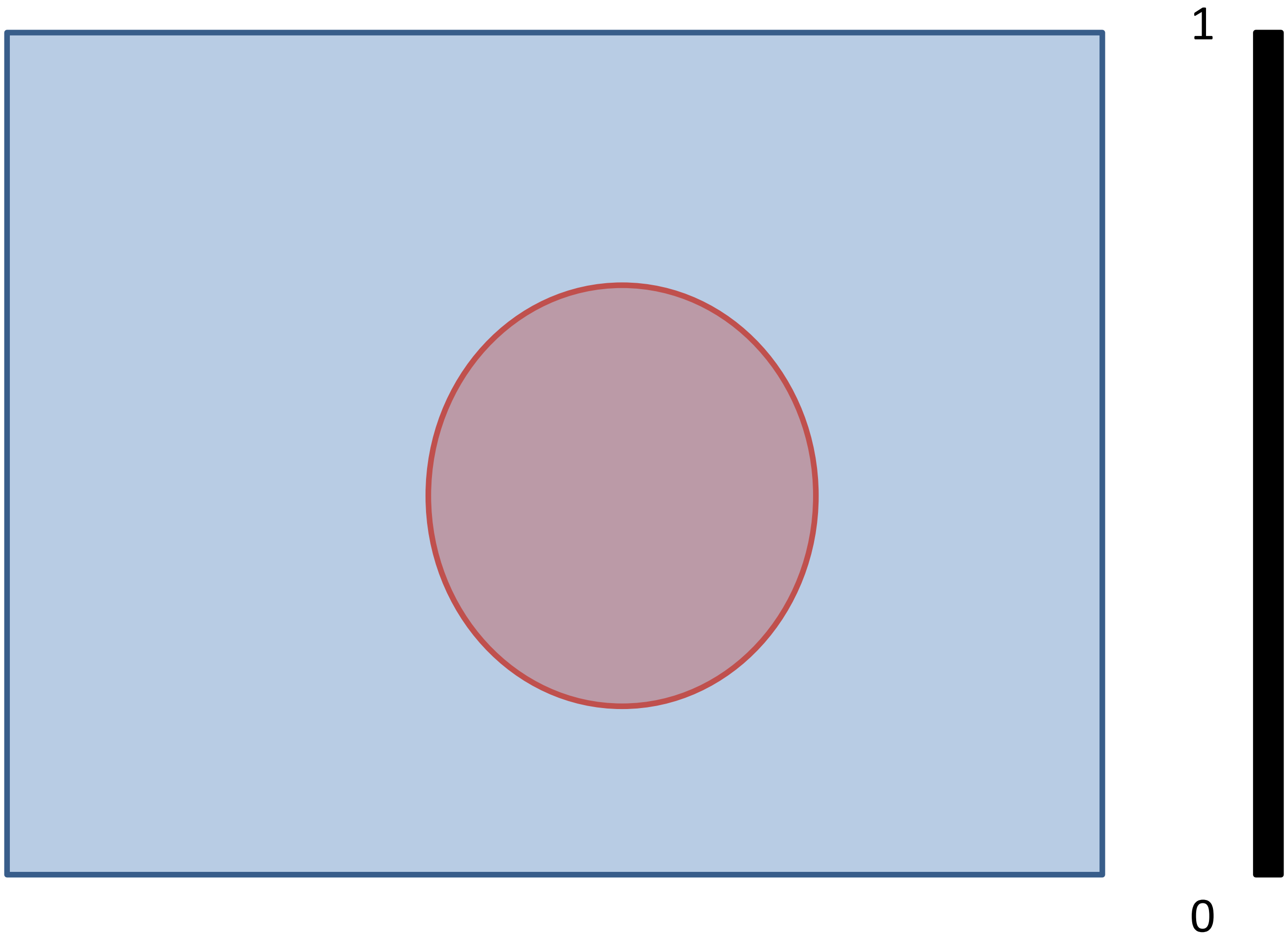
Symmetric: p(x,y) = p(y,x)

Form a table based of outcomes: sum across cells = 1

p(x,y)	Y=0	Y=1
X="cat"	.04	.32
X="dog"	.2	.04
X="bird"	.1	.1
X="human"	.1	.1



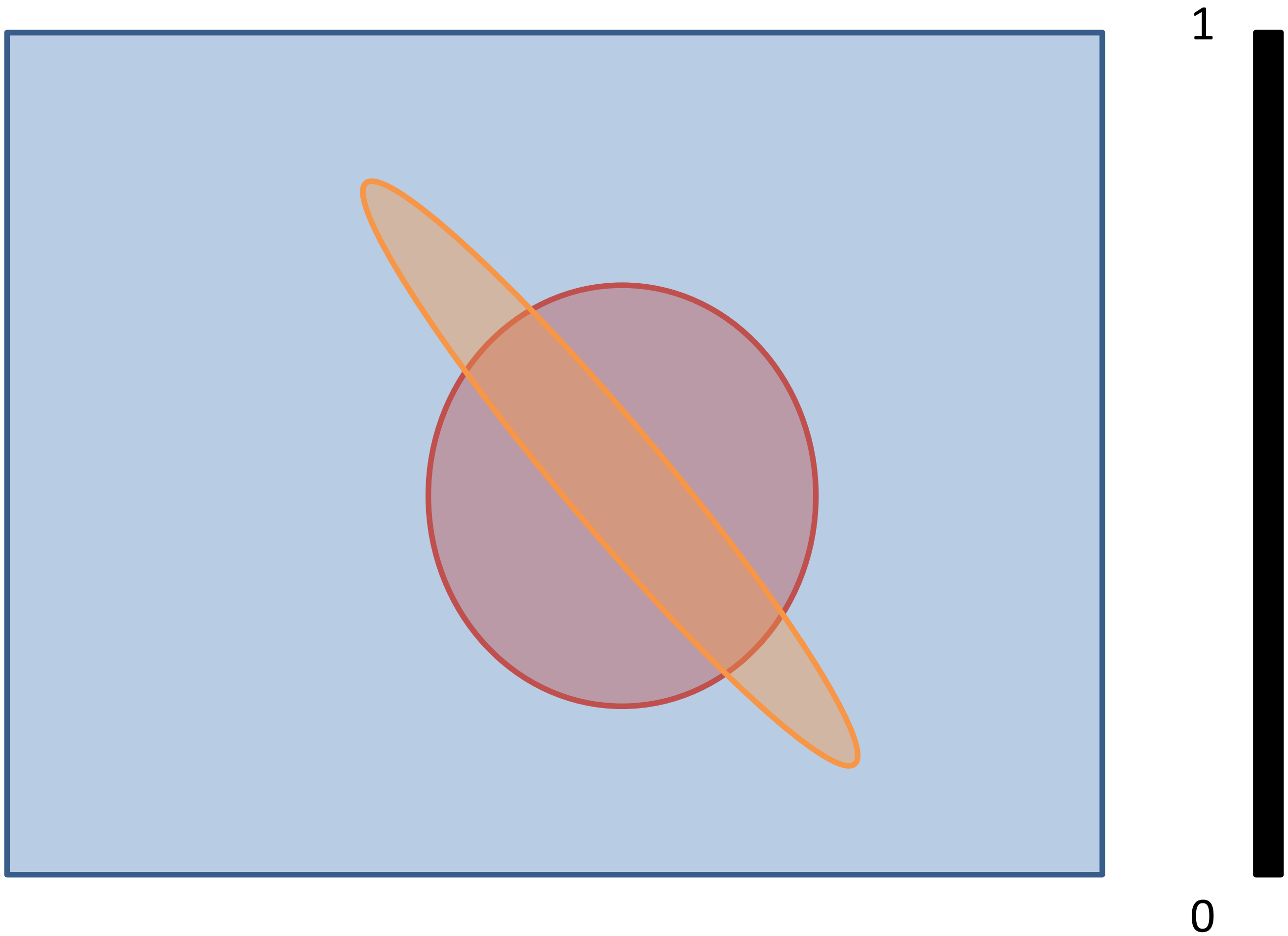
Joint Probabilities



p(A)

what happens as we add conjuncts?

Joint Probabilities

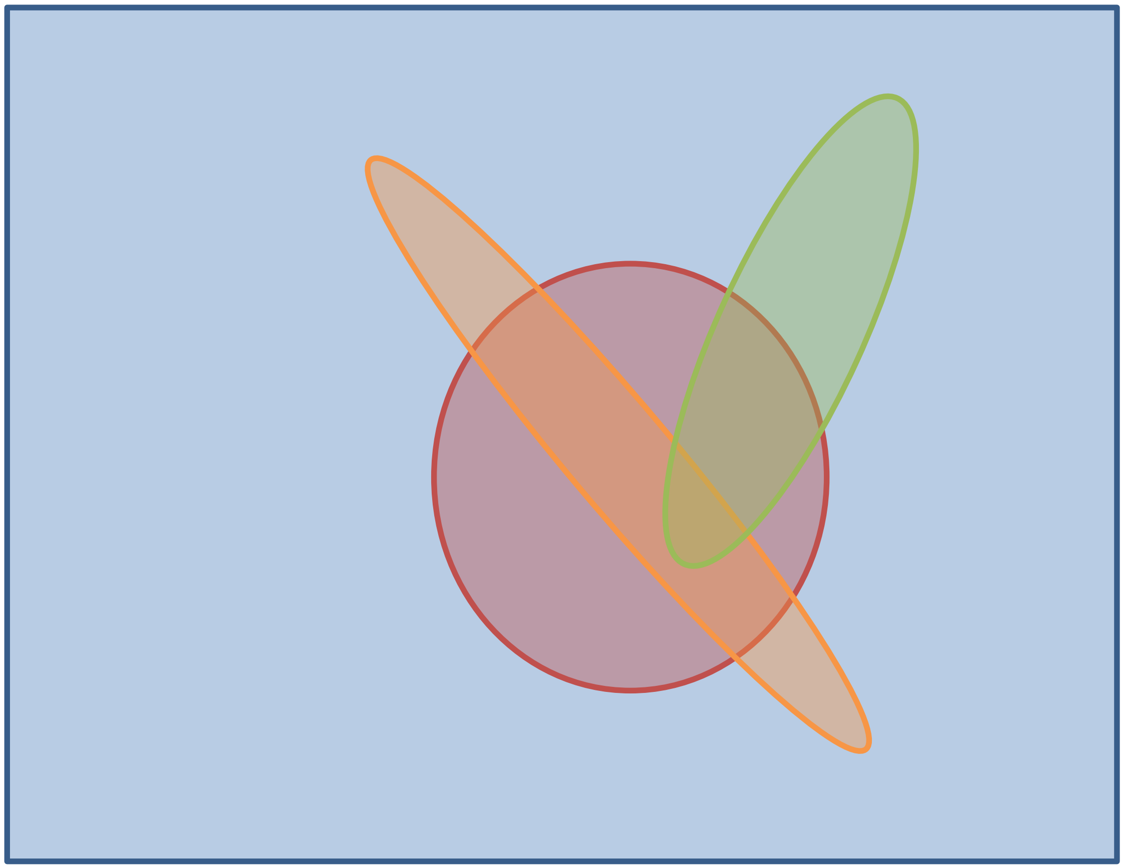


p(A)

p(A, B)

what happens as we add conjuncts?

Joint Probabilities



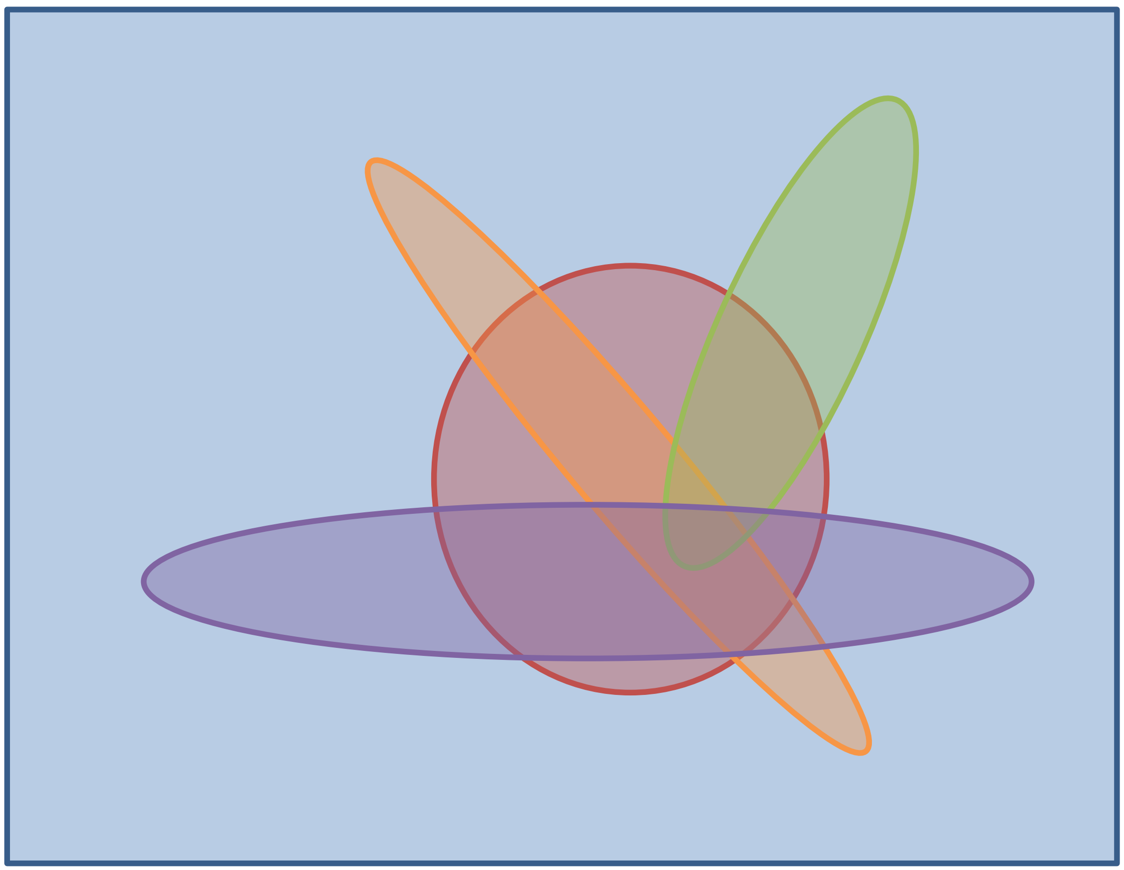
$p(A)$

$p(A, B)$

$p(A, B, C)$

what happens as we add conjuncts?

Joint Probabilities



1

0

p(A)

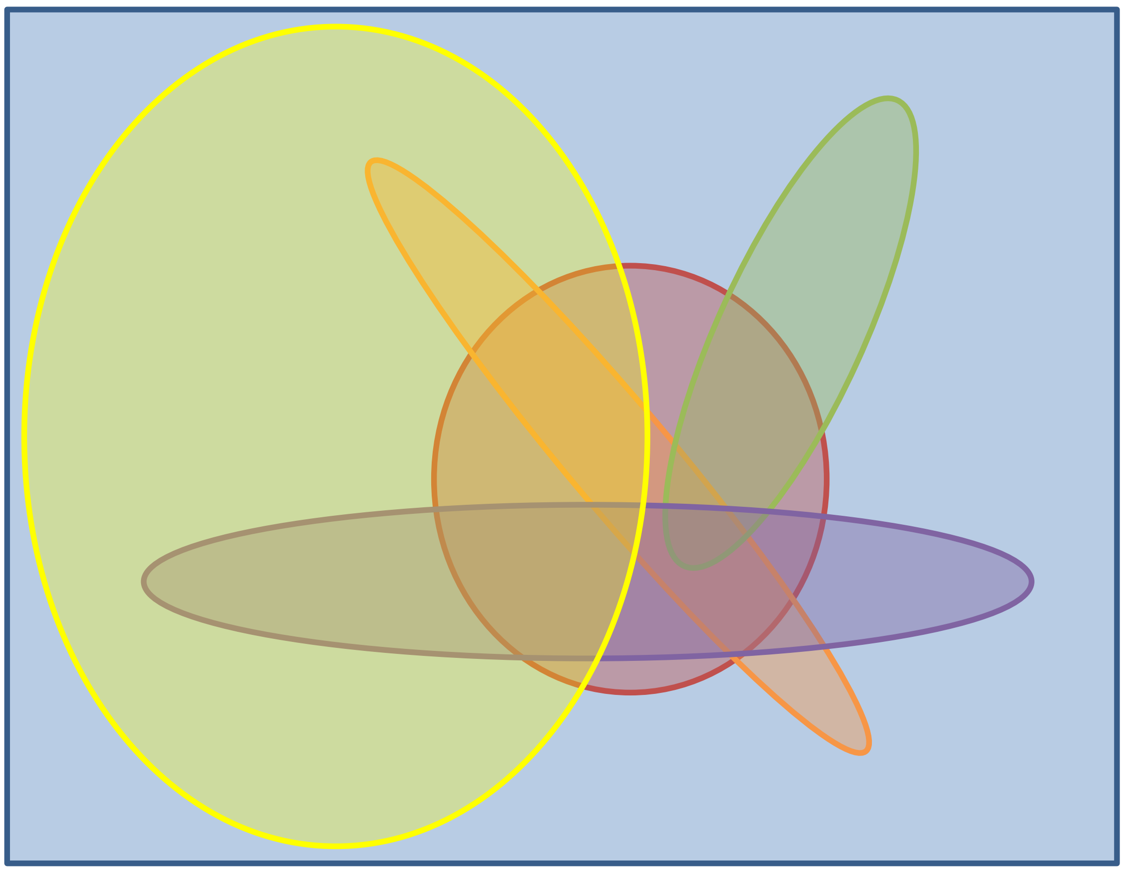
p(A, B)

p(A, B, C)

p(A, B, C, D)

what happens as we add conjuncts?

Joint Probabilities



1

p(A)

p(A, B)

p(A, B, C)

p(A, B, C, D)

p(A, B, C, D, E)

0

what happens as we add conjuncts?

Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable

Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x)*p(y)$

Generalizable to > 2 random variables

Q: Are the results of flipping the same coin twice in succession independent?

Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x)*p(y)$

Generalizable to > 2 random variables

Q: Are the results of flipping the same coin twice in succession independent?

A: Yes (assuming no weird effects)

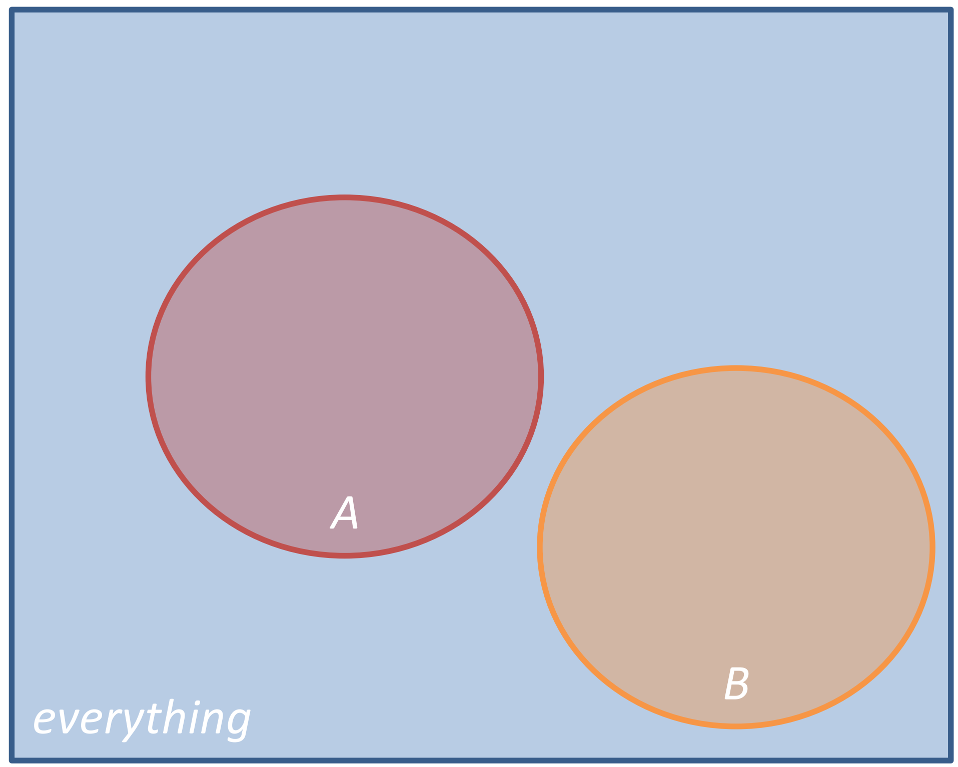
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: p(x,y) = p(x)*p(y)

Generalizable to > 2 random variables

Q: Are A and B independent?



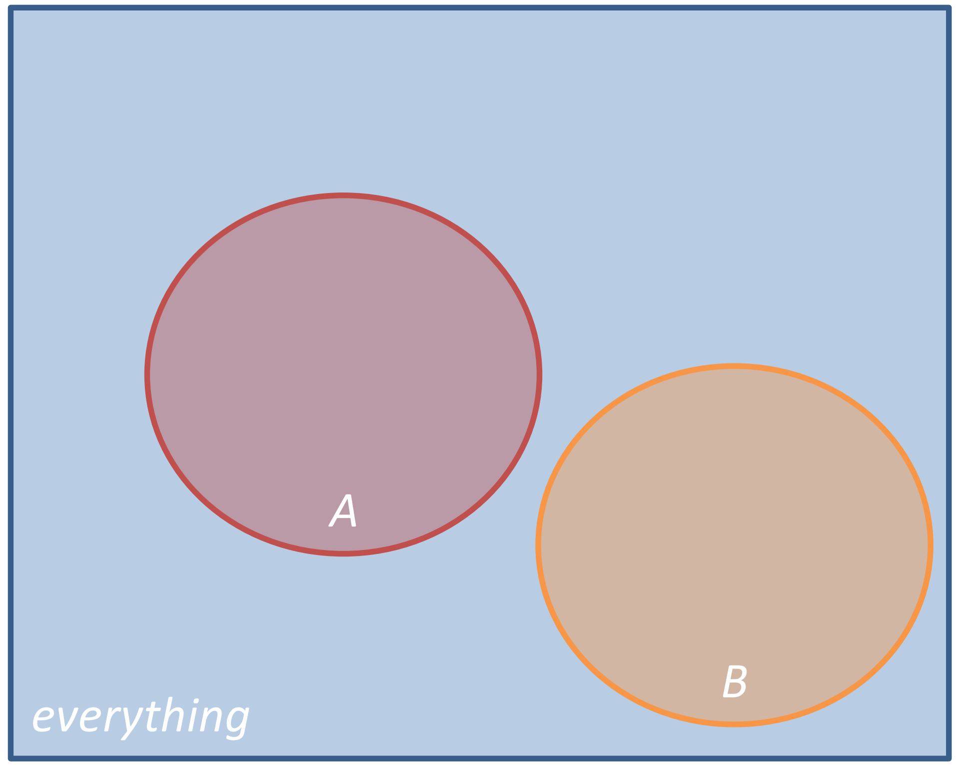
Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: $p(x,y) = p(x)*p(y)$

Generalizable to > 2 random variables

Q: Are A and B independent?



A: No (work it out from p(A,B)) and the axioms

Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: p(x,y) = p(x)*p(y)

Generalizable to > 2 random variables

Q: Are X and Y independent?

p(x,y)	Y=0	Y=1
X="cat"	.04	.32
X="dog"	.2	.04
X="bird"	.1	.1
X="human"	.1	.1

Probabilistic Independence

Independence: when events can occur and not impact the probability of other events

Formally: p(x,y) = p(x)*p(y)

Generalizable to > 2 random variables

Q: Are X and Y independent?

p(x,y)	Y=0	Y=1
X="cat"	.04	.32
X="dog"	.2	.04
X="bird"	.1	.1
X="human"	.1	.1

A: No (find the marginal probabilities of p(x) and p(y))

Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

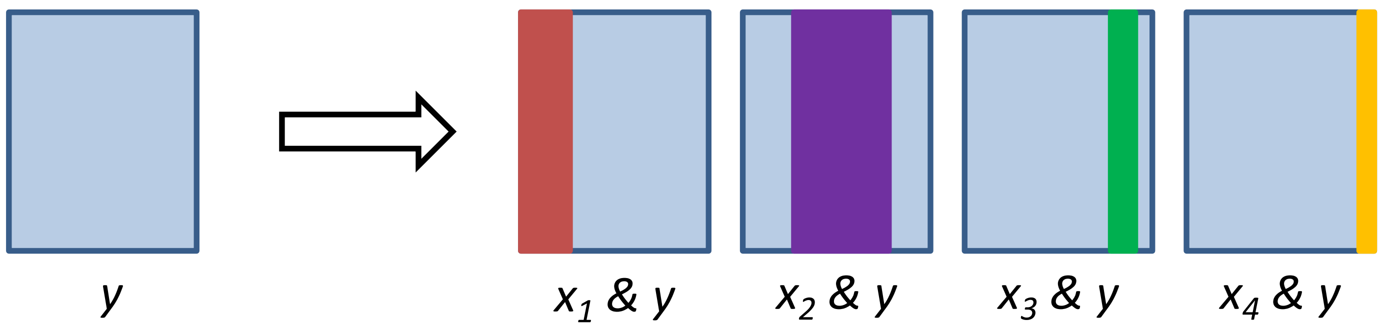
Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable

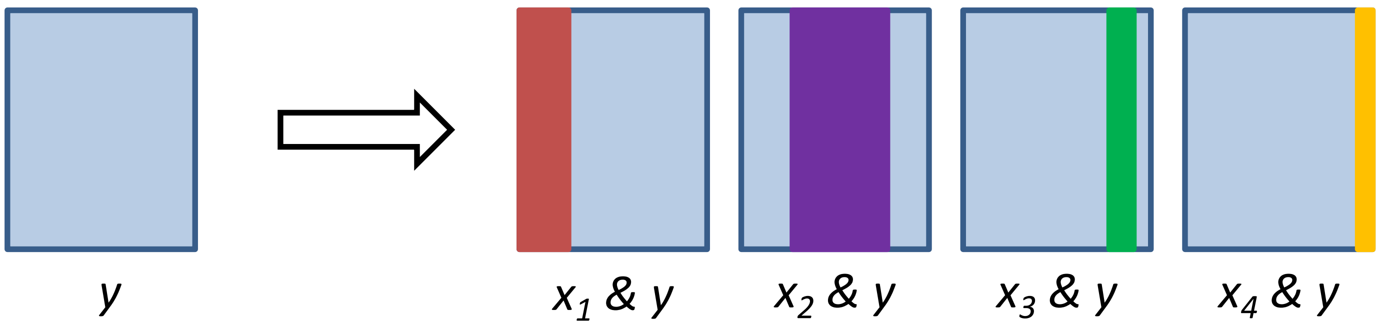
Marginal(ized) Probability: The Discrete Case



Consider the **mutually exclusive** ways that different values of x could occur with y

Q: How do write this in terms of joint probabilities?

Marginal(ized) Probability: The Discrete Case



Consider the **mutually exclusive** ways that different values of x could occur with y

$$p(y) = \sum_{x} p(x, y)$$

Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable

Conditional Probability

$$p(X \mid Y) = \frac{p(X, Y)}{p(Y)}$$

Conditional Probabilities
are Probabilities

Conditional Probability

$$p(X \mid Y) = \frac{p(X, Y)}{p(Y)}$$

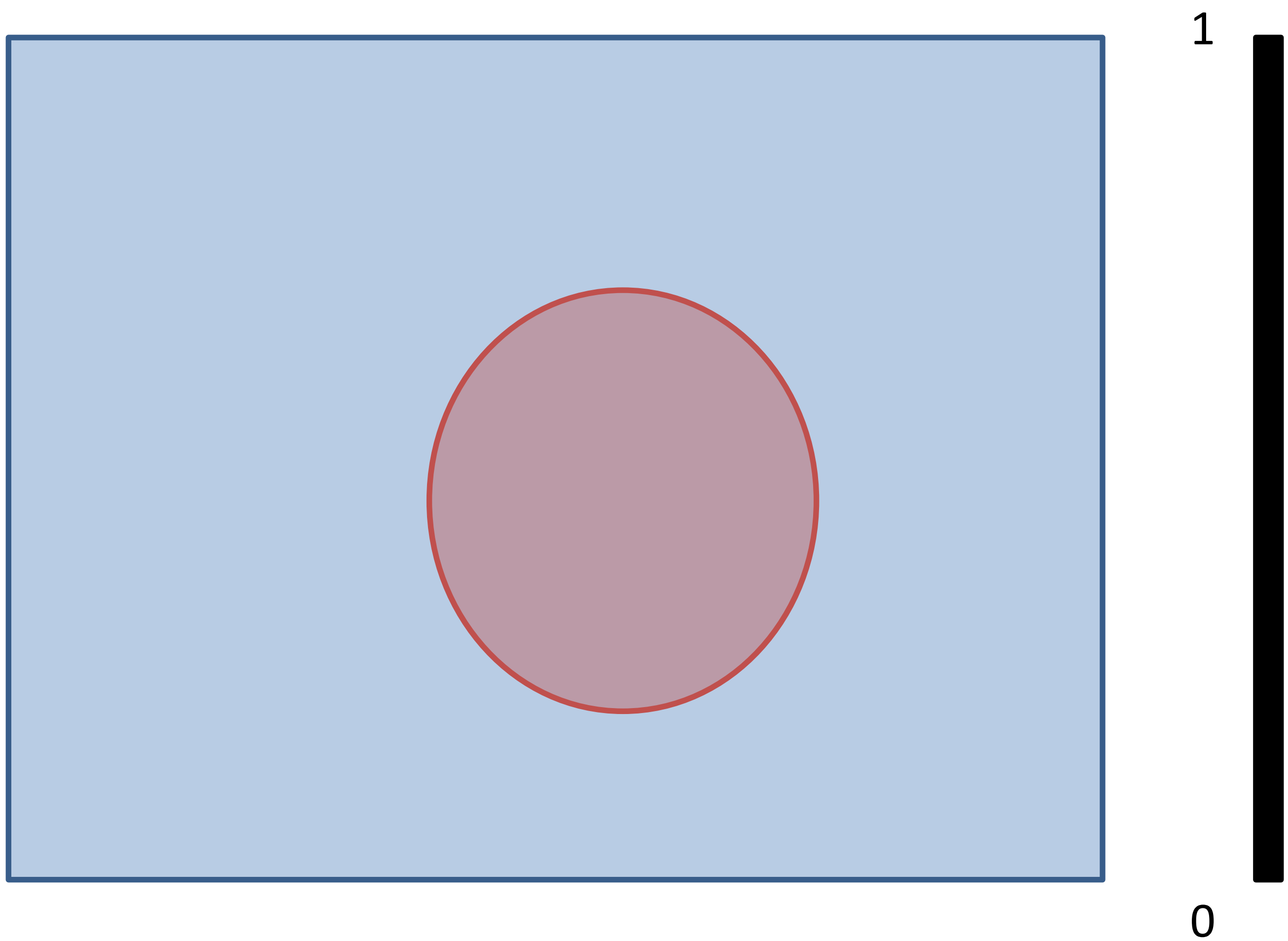
$p(Y) =$ marginal probability of Y

Conditional Probability

$$p(X \mid Y) = \frac{p(X, Y)}{p(Y)}$$

$$p(Y) = \int p(X, Y) dX$$

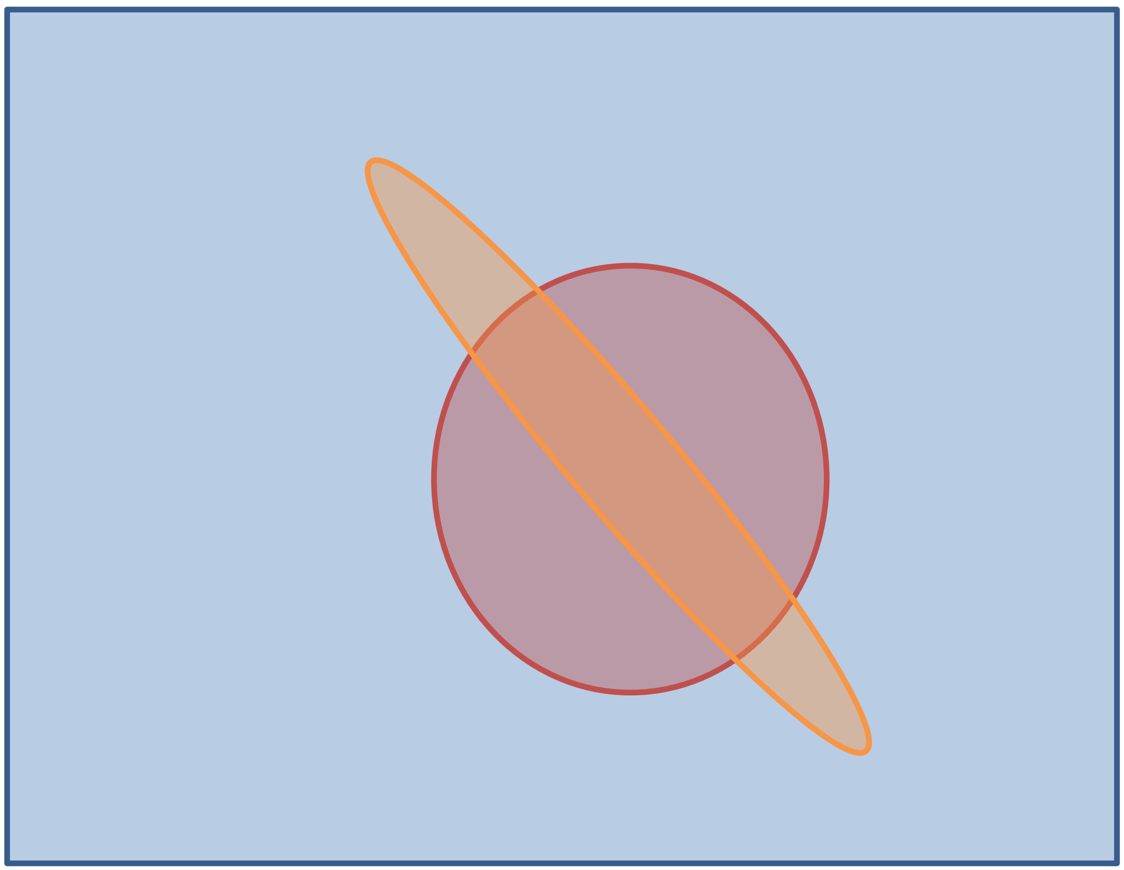
Conditional Probabilities: Changing the Right



p(A)

what happens as we add conjuncts to the right?

Conditional Probabilities: Changing the Right

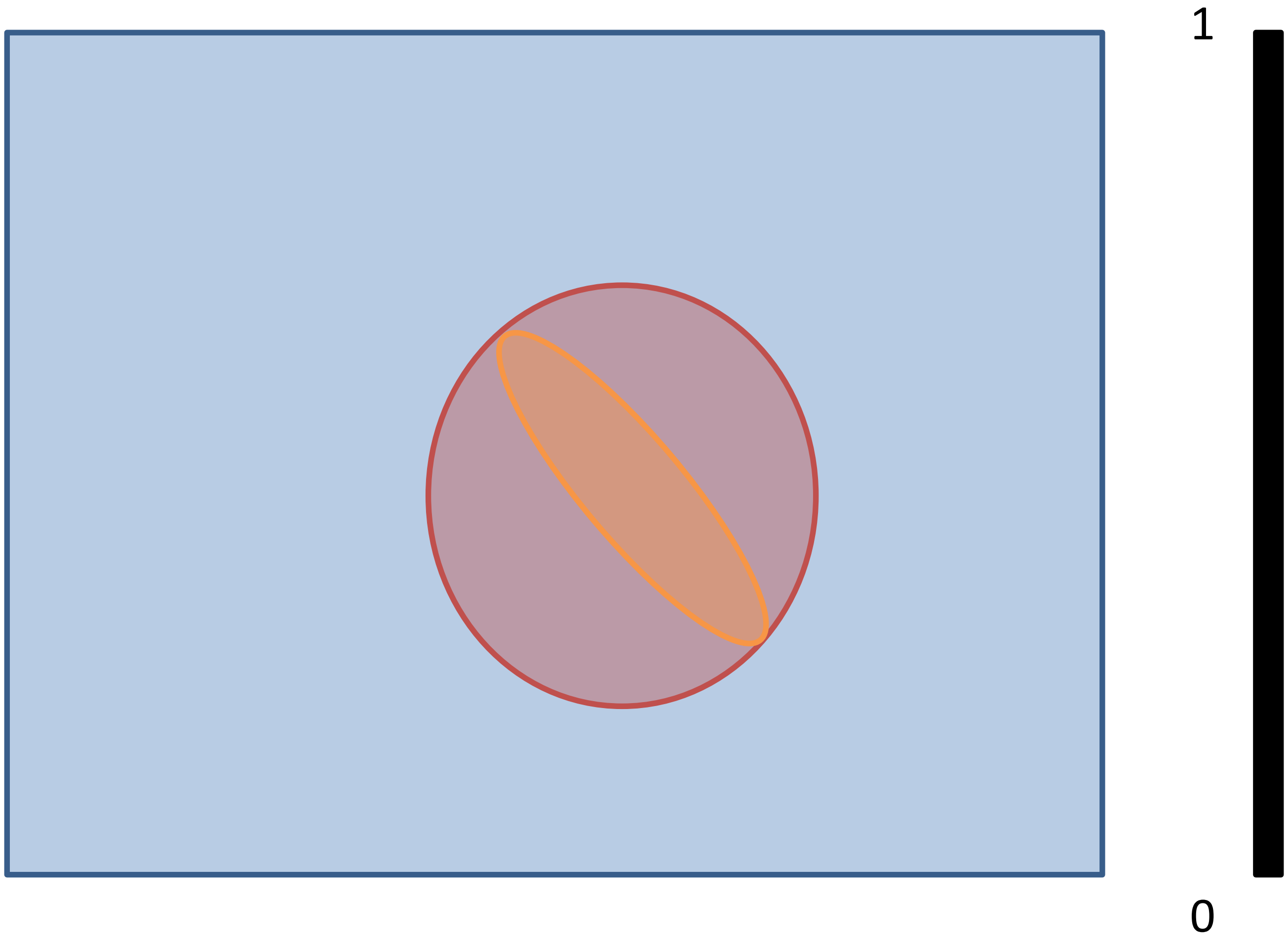


p(A)

p(A | B)

what happens as we add conjuncts to the right?

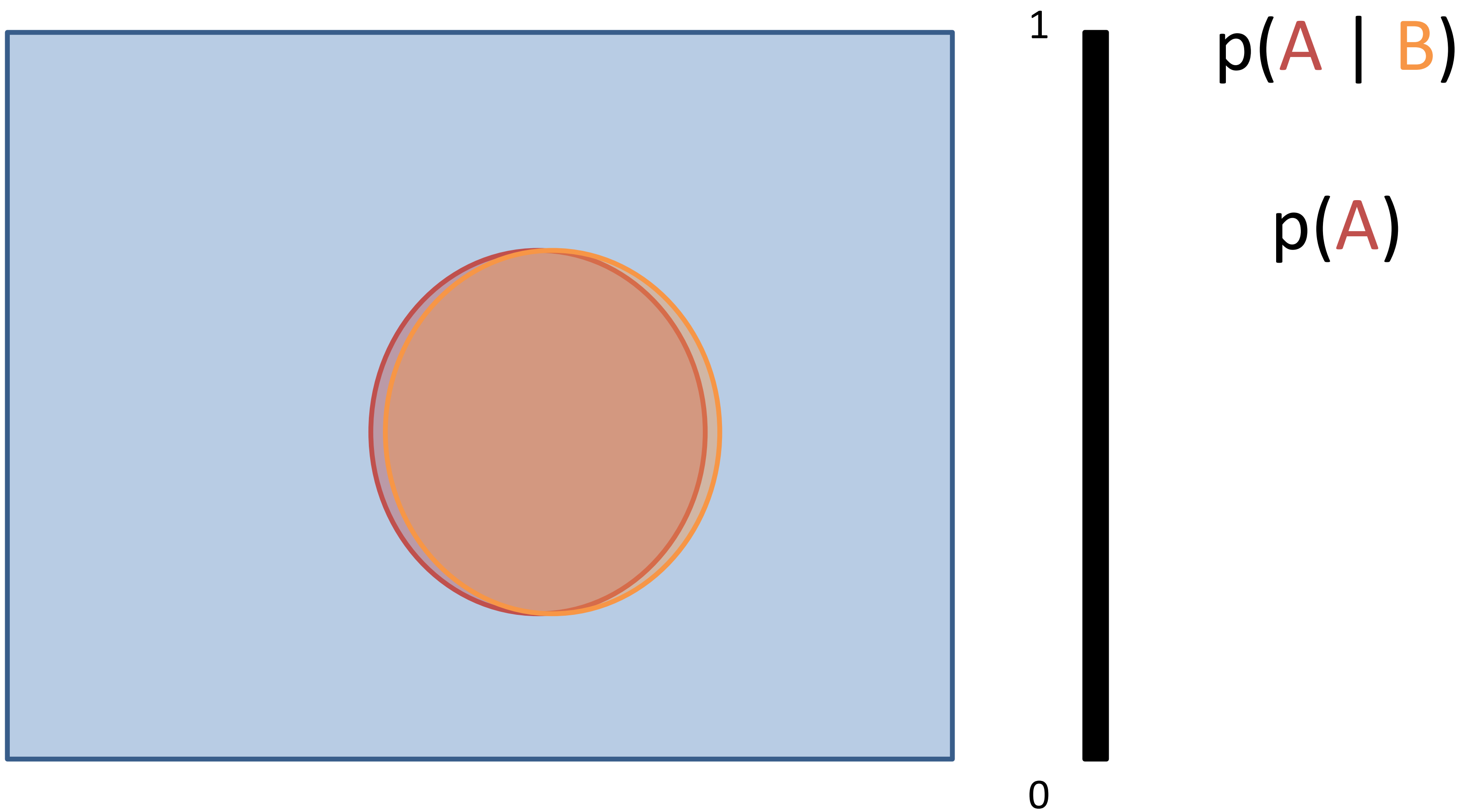
Conditional Probabilities: Changing the Right



$p(A \mid B)$
$p(A)$

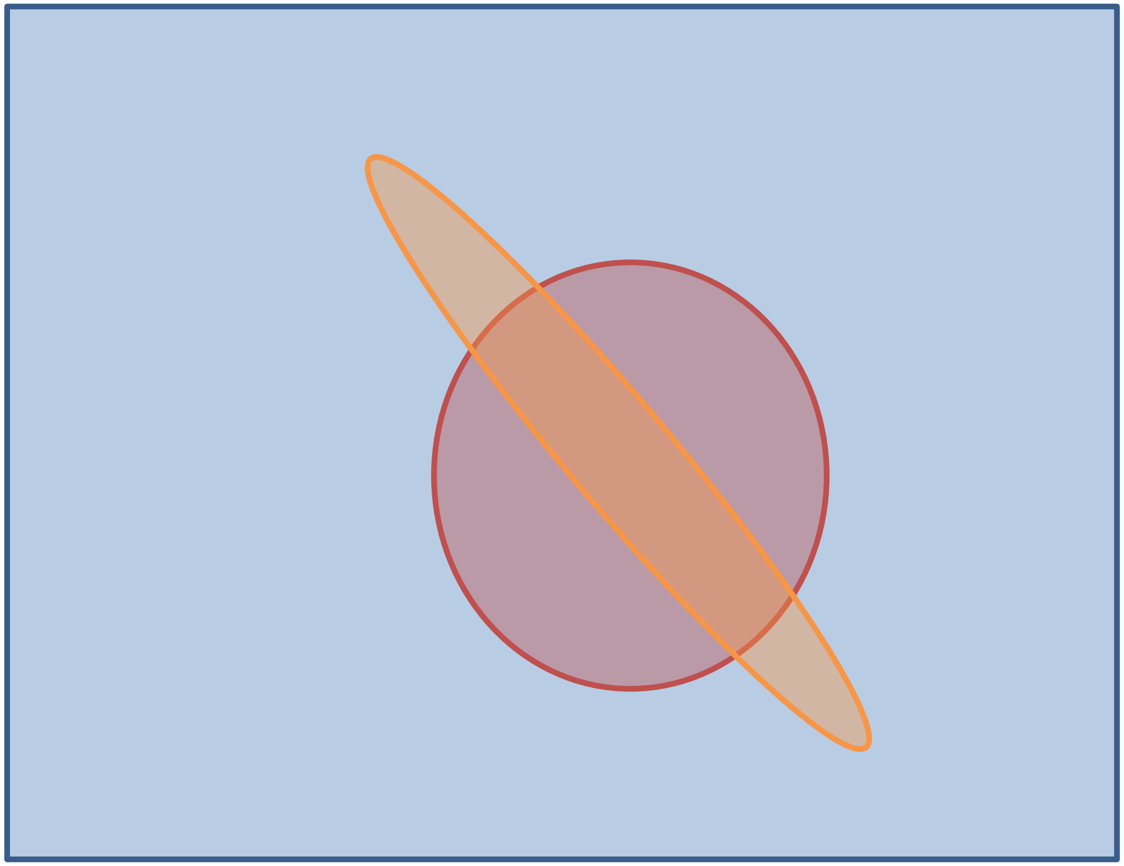
what happens as we add conjuncts to the right?

Conditional Probabilities: Changing the Right



what happens as we add conjuncts to the right?

Conditional Probabilities

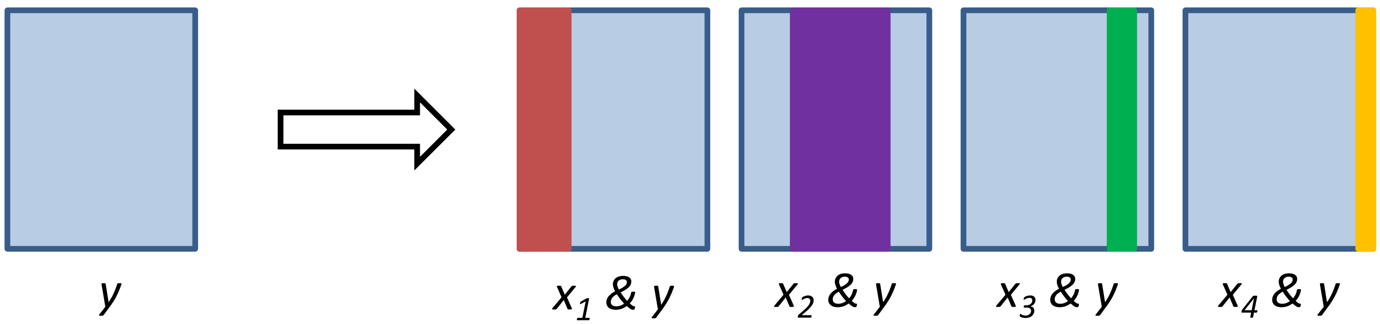


Bias vs. Variance

Lower bias: More specific to what we care about

Higher variance: For fixed observations, estimates become less reliable

Revisiting Marginal Probability: The Discrete Case



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

$$= \sum_x p(x)p(y \mid x)$$

Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable

Deriving Bayes Rule

Start with conditional
$p(X \mid Y)$

Deriving Bayes Rule

$$p(X \mid Y) = \frac{p(X, Y)}{p(Y)}$$

Solve for p(x,y)

Deriving Bayes Rule

$$p(X \mid Y) = \frac{p(X,Y)}{p(Y)}$$

Solve for p(x,y)

$$p(X,Y) = p(X \mid Y)p(Y)$$

p(x,y) = p(y,x)

$$p(X \mid Y) = \frac{p(Y \mid X) * p(X)}{p(Y)}$$

Bayes Rule

prior probability

likelihood

$$p(X \mid Y) = \frac{p(Y \mid X) * p(X)}{p(Y)}$$

posterior probability

marginal likelihood (probability)

Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable

Probability Chain Rule

$$p(x_1, x_2, \ldots, x_S) =$$
$$p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2) \cdots p(x_S \mid x_1, \ldots, x_i) =$$
$$\prod_{i}^{S} p(x_i \mid x_1, \ldots, x_{i-1})$$

extension of Bayes rule

Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable

Distribution Notation

If X is a R.V. and G is a distribution:

- $X \sim G$ means X is distributed according to ("sampled from") G

Distribution Notation

If X is a R.V. and G is a distribution:

- $X \sim G$ means X is distributed according to ("sampled from") G
- G often has parameters $\rho = (\rho_1, \rho_2, \ldots, \rho_M)$ that govern its "shape"
- Formally written as $X \sim G(\rho)$

Distribution Notation

If X is a R.V. and G is a distribution:

- $X \sim G$ means X is distributed according to ("sampled from") G
- G often has parameters $\rho = (\rho_1, \rho_2, \ldots, \rho_M)$ that govern its "shape"
- Formally written as $X \sim G(\rho)$

i.i.d. If $X_1, X_2, \ldots, X_{\mathrm{N}}$ are all independently sampled from $G(\rho)$, they are **i**ndependently and **i**dentically **d**istributed

Common Distributions

Bernoulli/Binomial

Categorical/Multinomial

Poisson

Normal

(Gamma)

Bernoulli: A single draw

- Binary R.V.: 0 (failure) or 1 (success)
- $X \sim \text{Bernoulli}(\rho)$
- $p(X = 1) = \rho, p(X = 0) = 1 - \rho$
- Generally, $p(X = k) = \rho^k (1 - p)^{1-k}$

Common Distributions

Bernoulli/Binomial

Categorical/Multinomial

Poisson

Normal

(Gamma)

Bernoulli: A single draw

- Binary R.V.: 0 (failure) or 1 (success)
- $X \sim \text{Bernoulli}(\rho)$
- $p(X = 1) = \rho, p(X = 0) = 1 - \rho$
- Generally, $p(X = k) = \rho^k (1 - p)^{1-k}$

Binomial: Sum of N iid Bernoulli draws

- Values X can take: 0, 1, …, N
- Represents number of successes
- $X \sim \text{Binomial}(N, \rho)$
- $p(X = k) = \binom{N}{k} \rho^k (1 - \rho)^{N-k}$

Common Distributions

Bernoulli/Binomial

Categorical/Multinomial

Poisson

Normal

(Gamma)

Categorical: A single draw

- Finite R.V. taking one of K values: 1, 2, …, K
- $X \sim \text{Cat}(\rho), \rho \in \mathbb{R}^K$
- $p(X = 1) = \rho_1, p(X = 2) = \rho_2, \dots p(X = K) = \rho_K$
- Generally, $p(X = k) = \prod_j \rho_j^{\mathbf{1}[k=j]}$
- $1[c] = \begin{cases} 1, & c \text{ is true} \\ 0, & c \text{ is false} \end{cases}$

Multinomial: Sum of N iid Categorical draws

- Vector of size K representing how often value k was drawn
- $X \sim \text{Multinomial}(N, \rho), \rho \in \mathbb{R}^K$

Common Distributions

Bernoulli/Binomial

Categorical/Multinomial

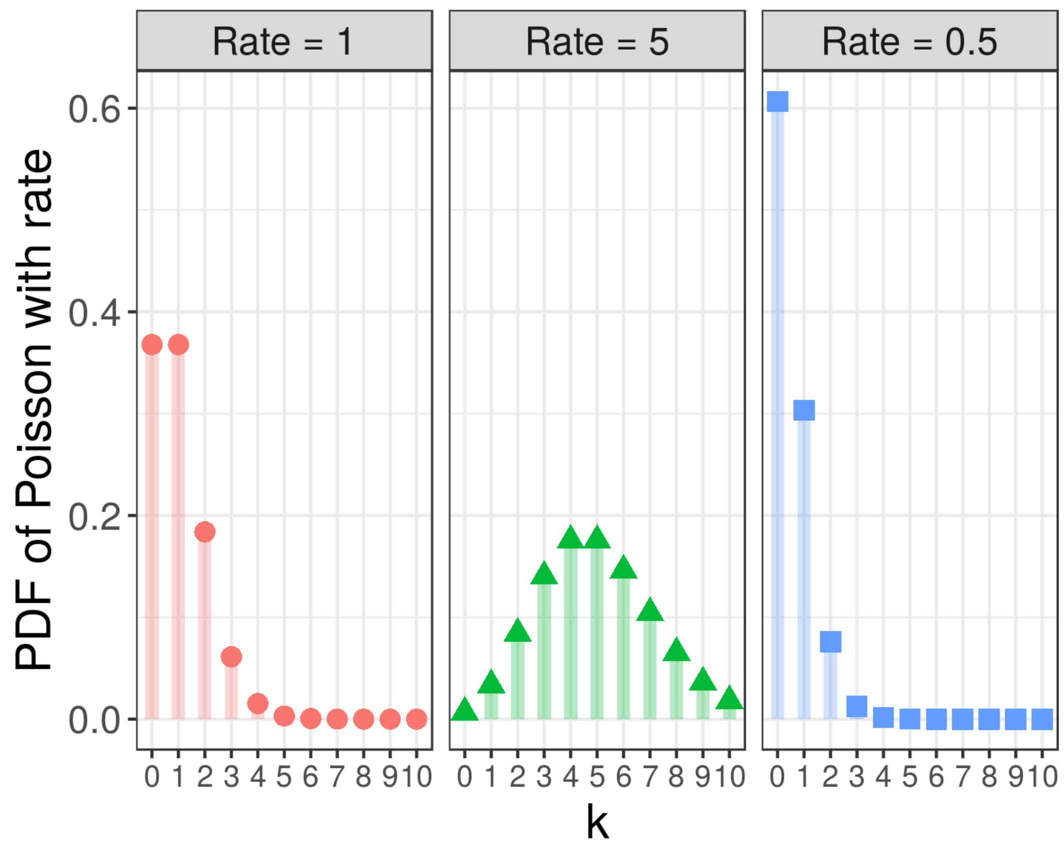
Poisson

Normal

(Gamma)

Poisson

- Finite R.V. taking any integer that is >= 0
- $X \sim \text{Poisson}(\lambda), \lambda \in \mathbb{R}$ is the "rate"
- $p(X = k) = \dfrac{\lambda^k \exp(-\lambda)}{k!}$



Common Distributions

Bernoulli/Binomial

Categorical/Multinomial

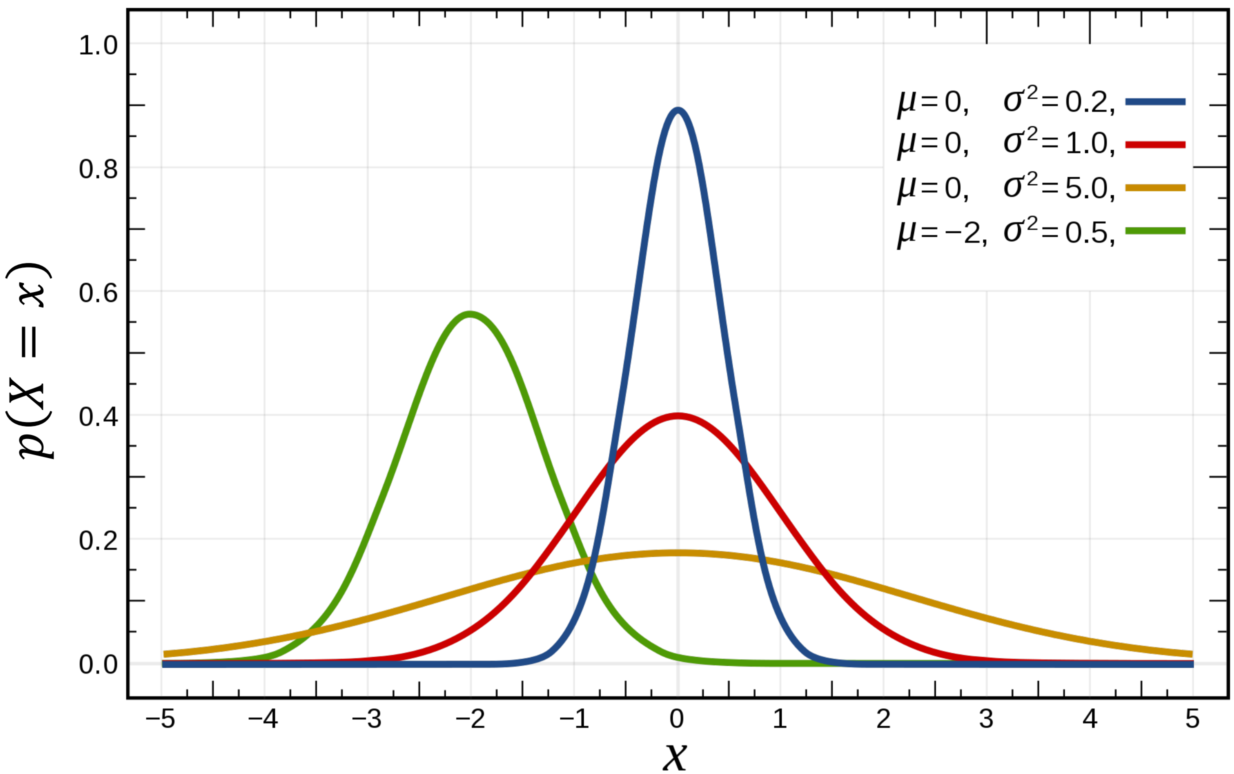
Poisson

Normal

(Gamma)

Normal

- Real R.V. taking any real number
- $X \sim \text{Normal}(\mu, \sigma)$, μ is the mean, σ is the standard deviation
- $p(X = x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$



https://upload.wikimedia.org/wikipedia/commons/thumb/7/74/Normal_Distribution_PDF.svg/1920px-Normal_Distribution_PDF.svg.png

Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable

Expected Value of a Random Variable

random variable

$$X \sim p(\cdot)$$

Expected Value of a Random Variable

random variable

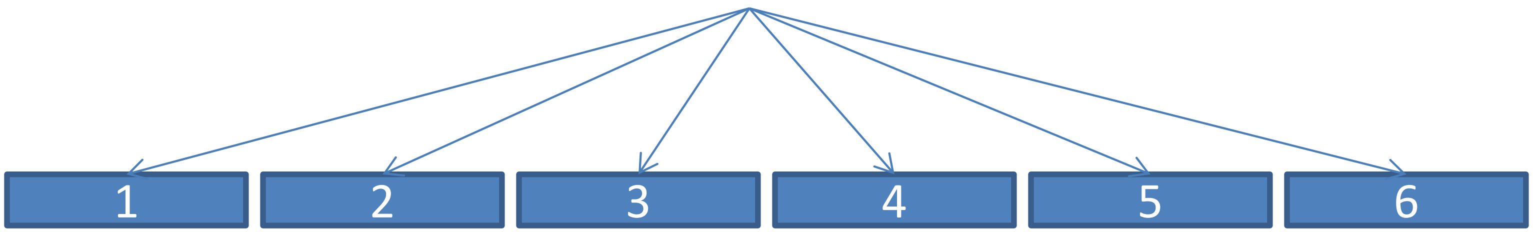
$$X \sim p(\cdot)$$

$$\mathbb{E}[X] = \sum_{x} x \, p(x)$$

expected value (distribution p is implicit)

Expected Value: Example

uniform distribution of number of cats I have

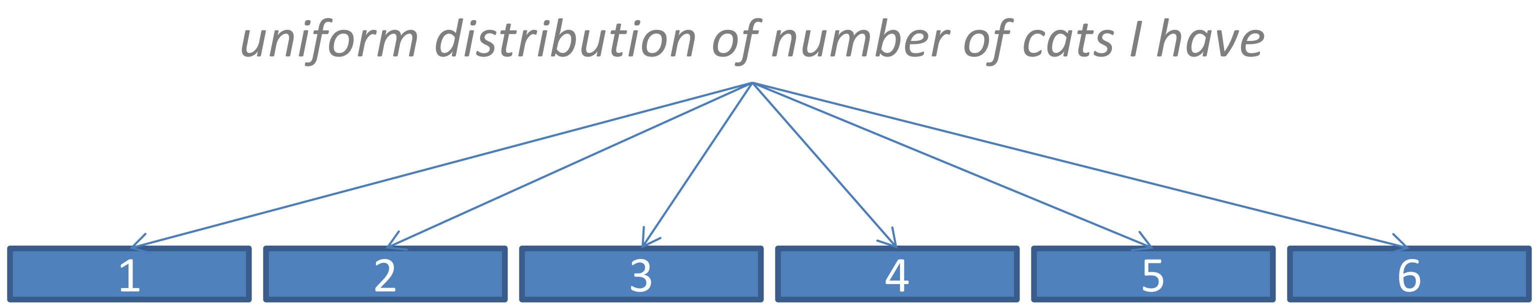


$$\mathbb{E}[X] = \sum_x x \, p(x)$$

1/6 * 1 +
1/6 * 2 +
1/6 * 3 +
1/6 * 4 +
1/6 * 5 +
1/6 * 6

= 3.5

Expected Value: Example



$$\mathbb{E}[X] = \sum_{x} x\, p(x)$$

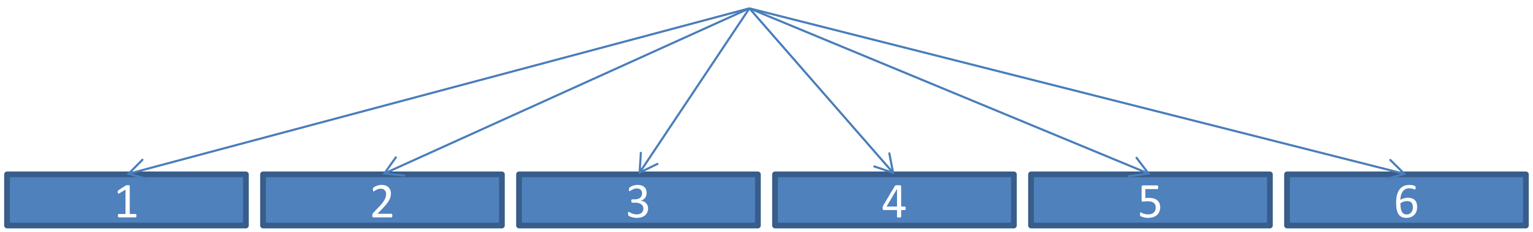
1/6 * 1 +
1/6 * 2 +
1/6 * 3 +
1/6 * 4 +
1/6 * 5 +
1/6 * 6

= 3.5

Q: What common distribution is this?

Expected Value: Example

uniform distribution of number of cats I have



$$\mathbb{E}[X] = \sum_x x \, p(x)$$

1/6 * 1 +
1/6 * 2 +
1/6 * 3 +
1/6 * 4 +
1/6 * 5 +
1/6 * 6

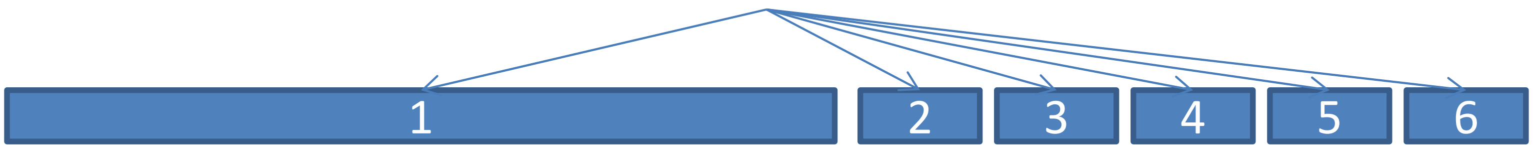
= 3.5

Q: What common distribution is this?

A: Categorical

Expected Value: Example 2

non-uniform distribution of number of cats a normal cat person has



$$\mathbb{E}[X] = \sum_x x\, p(x)$$

1/2 * 1 +
1/10 * 2 +
1/10 * 3 +
1/10 * 4 +
1/10 * 5 +
1/10 * 6

= 2.5

Expected Value of a Function of a Random Variable

$$X \sim p(\cdot)$$

$$\mathbb{E}[X] = \sum_x x\, p(x)$$

$$\mathbb{E}[f(X)] = ???$$

Expected Value of a Function of a Random Variable

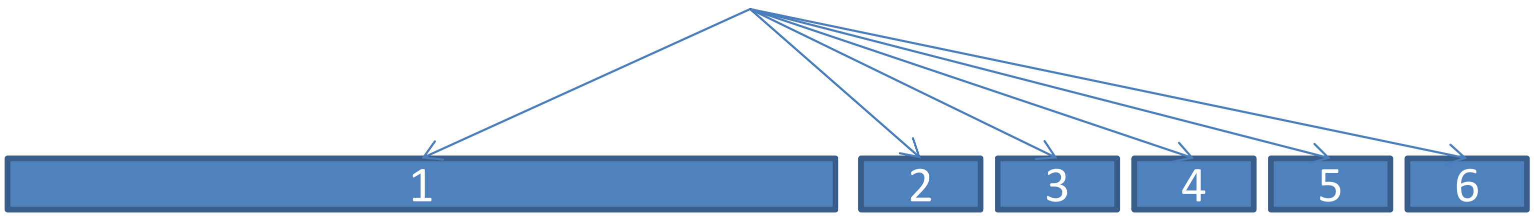
$$X \sim p(\cdot)$$

$$\mathbb{E}[X] = \sum_x x\, p(x)$$

$$\mathbb{E}[f(X)] = \sum_x f(x)\, p(x)$$

Expected Value of Function: Example

non-uniform distribution of number of cats I start with



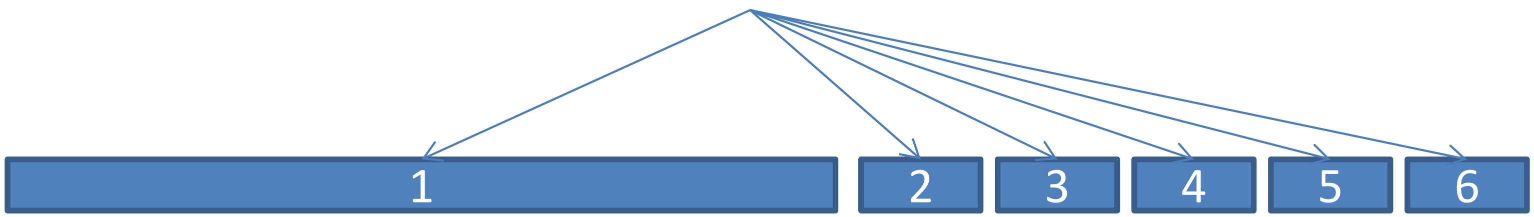
What if each cat magically becomes two?

$$f(k) = 2^k$$

$$\mathbb{E}[f(X)] = \sum_x f(x)\, p(x)$$

Expected Value of Function: Example

non-uniform distribution of number of cats I start with



What if each cat magically becomes two?

$$f(k) = 2^k$$

$$\mathbb{E}[f(X)] = \sum_x f(x)\, p(x) = \sum_x 2^x p(x)$$

$$\begin{array}{c} 1/2 * 2^1 + \\ 1/10 * 2^2 + \\ 1/10 * 2^3 + \\ 1/10 * 2^4 + \\ 1/10 * 2^5 + \\ 1/10 * 2^6 \end{array} \quad = 13.4$$

Probability Prerequisites

Basic probability axioms and definitions

Joint probability

Probabilistic Independence

Marginal probability

Definition of conditional probability

Bayes rule

Probability chain rule

Common distributions

Expected Value (of a function) of a Random Variable

Outline

Review+Extension

Probability

Decision Theory

Loss Functions

Decision Theory

"Decision theory is trivial, apart from the computational details" – MacKay, ITILA, Ch 36

Input: **x** ("state of the world")

Output: a decision $\hat{y}$

Decision Theory

"Decision theory is trivial, apart from the computational details" – MacKay, ITILA, Ch 36

Input: $\mathbf{x}$ ("state of the world")

Output: a decision $\hat{y}$

Requirement 1: a decision (hypothesis) function $h(\mathbf{x})$ to produce $\hat{y}$

Decision Theory

"Decision theory is trivial, apart from the computational details" – MacKay, ITILA, Ch 36

Input: **x** ("state of the world")

Output: a decision ŷ

Requirement 1: a decision (hypothesis) function h(**x**) to produce ŷ

Requirement 2: a function ℓ(y, ŷ) telling us how wrong we are

Decision Theory

"Decision theory is trivial, apart from the computational details" – MacKay, ITILA, Ch 36

Input: **x** ("state of the world")

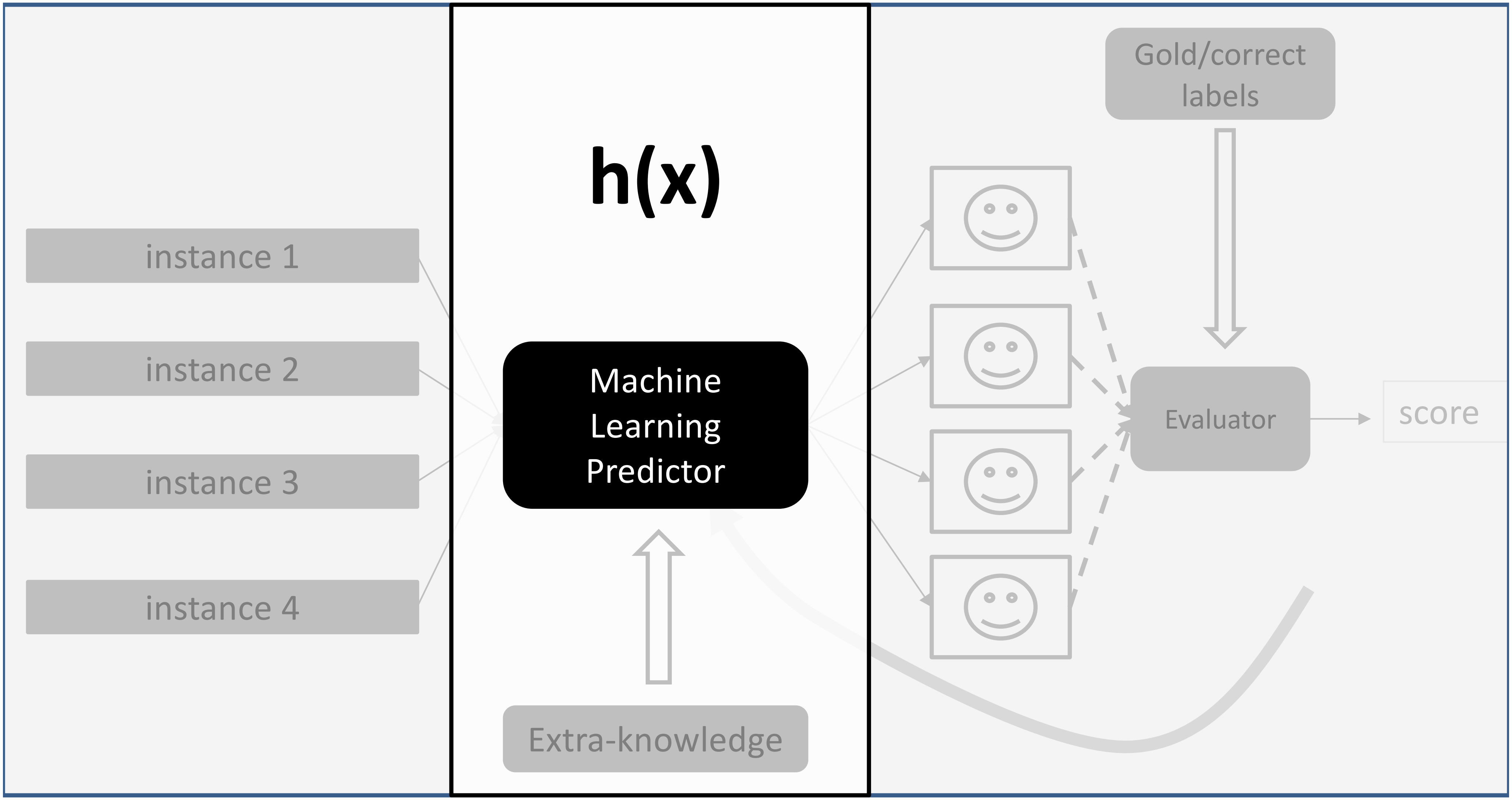
Output: a decision ŷ

Requirement 1: a decision (hypothesis) function h(**x**) to produce ŷ

Requirement 2: a loss function ℓ(y, ŷ) telling us how wrong we are

Goal: minimize our *expected* loss across any possible input

Requirement 1: Decision Function



h(x) is our predictor (classifier, regression model, clustering model, etc.)

Requirement 2: Loss Function

"ell" (fancy l character)

predicted label/result

$$\ell(y, \hat{y}) \geq 0$$

"correct" label/result

optimize ℓ?

- minimize
- ~~maximize~~

loss: A function that tells you how much to penalize a prediction $\hat{y}$ from the correct answer y

Requirement 2: Loss Function

"ell" (fancy l character)

predicted label/result

Negative ℓ ($-\ell$) is called a *utility* or *reward* function

$$\ell(y, \hat{y}) \geq 0$$

"correct" label/result

loss: A function that tells you how much to penalize a prediction $\hat{y}$ from the correct answer y

Decision Theory

minimize expected loss across any possible input

$$\arg\min_{\hat{y}} \mathbb{E}[\ell(y, \hat{y})]$$

Risk Minimization

minimize expected loss across any possible input

$$\arg\min_{\hat{y}} \mathbb{E}[\ell(y, \hat{y})] = \arg\min_{h} \mathbb{E}[\ell(y, h(\boldsymbol{x}))]$$

a *particular*, unspecified input pair (**x**,y)… but we want any possible pair

Decision Theory

minimize expected loss across any possible input

$$\arg\min_{\hat{y}} \mathbb{E}[\ell(y, \hat{y})] =$$

$$\arg\min_{h} \mathbb{E}[\ell(y, h(\boldsymbol{x}))] =$$

$$\operatorname*{argmin}_{\mathrm{h}} \mathbb{E}_{(\boldsymbol{x}, y) \sim P}\left[\ell\left(y, h(\boldsymbol{x})\right)\right]$$

Assumption: there exists *some* true (but likely unknown) distribution *P* over inputs **x** and outputs y

Risk Minimization

minimize expected loss across any possible input

$$\arg\min_{\hat{y}} \mathbb{E}[\ell(y, \hat{y})] =$$

$$\arg\min_{h} \mathbb{E}[\ell(y, h(\boldsymbol{x}))] =$$

$$\underset{\mathrm{h}}{\operatorname{argmin}}\ \mathbb{E}_{(\boldsymbol{x},y)\sim P}\left[\ell\big(y, h(\boldsymbol{x})\big)\right] =$$

$$\underset{\mathrm{h}}{\operatorname{argmin}} \int \ell\big(y, h(\boldsymbol{x})\big) P(\boldsymbol{x}, y)\, d(\boldsymbol{x}, y)$$

Risk Minimization

minimize expected loss across any possible input

$$\arg\min_{\hat{y}} \mathbb{E}[\ell(y, \hat{y})] =$$

$$\arg\min_{h} \mathbb{E}[\ell(y, h(\boldsymbol{x}))] =$$

$$\operatorname*{argmin}_{\mathrm{h}} \mathbb{E}_{(\boldsymbol{x},y)\sim P}\left[\ell\left(y, h(\boldsymbol{x})\right)\right] =$$

$$\operatorname*{argmin}_{\mathrm{h}} \int \ell\left(y, h(\boldsymbol{x})\right) P(\boldsymbol{x}, y)\, d(\boldsymbol{x}, y)$$

we don't know this distribution*!

**we could try to approximate it analytically*

Empirical Risk Minimization

minimize expected loss across our observed input

$$\arg\min_{\hat{y}} \mathbb{E}[\ell(y, \hat{y})] =$$

$$\arg\min_{h} \mathbb{E}[\ell(y, h(\boldsymbol{x}))] =$$

$$\operatorname*{argmin}_{\mathrm{h}} \mathbb{E}_{(\boldsymbol{x}, y) \sim P}\left[\ell\left(y, h(\boldsymbol{x})\right)\right] \approx$$

$$\operatorname*{argmin}_{\mathrm{h}} \frac{1}{N} \sum_{i=1}^{N} \ell\left(y_i, h(\boldsymbol{x}_i)\right)$$

Empirical Risk Minimization

minimize expected loss across our observed input

$$\underset{h}{\operatorname{argmin}} \sum_{i=1}^{N} \ell\left(y_i, h(\boldsymbol{x}_i)\right)$$

our
classifier/predictor

controlled by our
parameters θ

change θ →
change the behavior of the classifier

Best Case: Optimize Empirical Risk with Gradients

$$\operatorname*{argmin}_{\mathrm{h}} \sum_{i=1}^{N} \ell\big(y_i, h_\theta(\boldsymbol{x}_i)\big)$$

change θ →
change the behavior of the classifier

$$\Downarrow$$

$$\operatorname*{argmin}_{\theta} \sum_{i=1}^{N} \ell\big(y_i, h_\theta(\boldsymbol{x}_i)\big)$$

Best Case: Optimize Empirical Risk with Gradients

$$\underbrace{\operatorname*{argmin}_{\theta} \sum_{i=1}^{N} \ell(y_i, h_\theta(\boldsymbol{x}_i))}_{F(\theta)}$$

change θ →
change the behavior of the classifier

How? Use Gradient Descent on $F(\theta)$!

differentiating might not always work: "… apart from the computational details"

Best Case: Optimize Empirical Risk with Gradients

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \ell\big(y_i, h_\theta(\boldsymbol{x}_i)\big)$$

change θ →
change the behavior of the classifier

$$\Downarrow$$

$$\nabla_\theta F = \sum_i \frac{\partial \ell\big(y_i, \hat{y} = h_\theta(\boldsymbol{x}_i)\big)}{\partial \hat{y}} \nabla_\theta h_\theta(\boldsymbol{x_i})$$

differentiating might not always work: "… apart from the computational details"

Best Case: Optimize Empirical Risk with Gradients

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \ell\left(y_i, h_\theta(\boldsymbol{x}_i)\right)$$

change θ →
change the behavior of the classifier

$$\nabla_\theta F = \sum_i \frac{\partial \ell\left(y_i, \hat{y} = h_\theta(\boldsymbol{x}_i)\right)}{\partial \hat{y}} \nabla_\theta h_\theta(\boldsymbol{x}_i)$$

Step 1: compute the gradient of the loss wrt the predicted value

differentiating might not always work: "… apart from the computational details"

Best Case: Optimize Empirical Risk with Gradients

$$\operatorname*{argmin}_{\theta} \sum_{i=1}^{N} \ell\big(y_i, h_\theta(\boldsymbol{x}_i)\big)$$

change θ →
change the behavior of the classifier

$$\Downarrow$$

$$\nabla_\theta F = \sum_i \frac{\partial \ell\big(y_i, \hat{y} = h_\theta(\boldsymbol{x}_i)\big)}{\partial \hat{y}} \nabla_\theta h_\theta(\boldsymbol{x}_i)$$

Step 1: compute the gradient of the loss wrt the predicted value

Step 2: compute the gradient of the predicted value wrt θ.

differentiating might not always work: "… apart from the computational details"

Outline

Review+Extension

Probability

Decision Theory

Loss Functions

Loss Functions Serve a Task

Classification Regression Clustering	Fully-supervised Semi-supervised Un-supervised	Probabilistic Generative Conditional Spectral Neural Memory-based Exemplar …
*the **task**: what* kind *of problem are you solving?*	*the **data**: amount of human input/number of labeled examples*	*the **approach**: how any data are being used*

Classification: Supervised Machine Learning

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- …

Input:

an instance d

a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

A training set of m hand-labeled instances $(d_1,c_1),\ldots,(d_m,c_m)$

Output:

a learned classifier γ that maps instances to classes

γ learns to associate certain ***features*** of instances with their labels

Classification Example: Face Recognition

Class	Image
Avrim	
Avrim	
Avrim	
Avrim	

Class	Image
Tom	
Tom	
Tom	
Tom	

Courtesy from Hamed Pirsiavash

Classification Loss Function Example: 0-1 Loss

$$\ell(y, \hat{y}) = \begin{cases} 0, & \text{if } y = \hat{y} \\ 1, & \text{if } y \neq \hat{y} \end{cases}$$

Classification Loss Function Example: 0-1 Loss

$$\ell(y, \hat{y}) = \begin{cases} 0, & \text{if } y = \hat{y} \\ 1, & \text{if } y \neq \hat{y} \end{cases}$$

Problem 1: not differentiable wrt $\hat{y}$ (or θ)

Classification Loss Function Example: 0-1 Loss

$$\ell(y, \hat{y}) = \begin{cases} 0, & \text{if } y = \hat{y} \\ 1, & \text{if } y \neq \hat{y} \end{cases}$$

Problem 1: not differentiable wrt $\hat{y}$ (or θ)

Solution 1: is the data linearly separable?
Perceptron (next class) can work

Classification Loss Function Example: 0-1 Loss

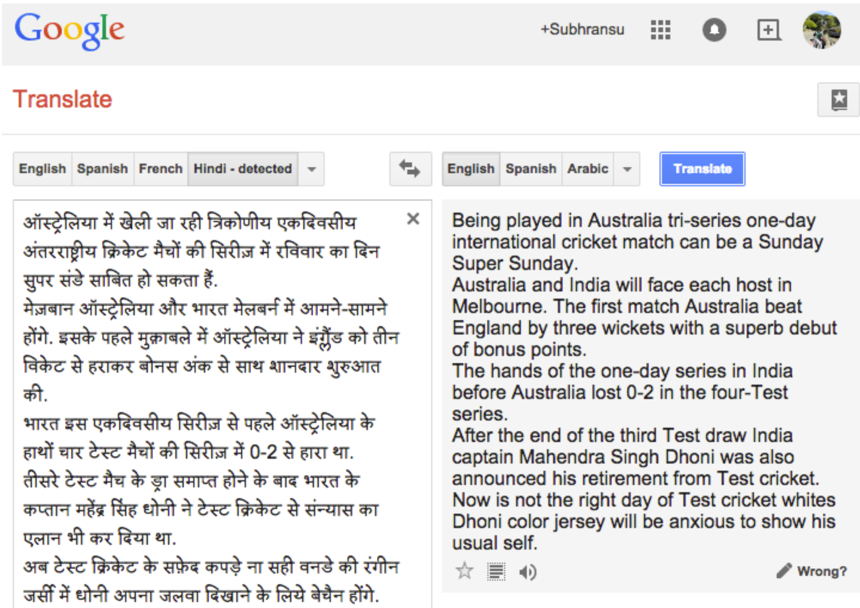
$$\ell(y, \hat{y}) = \begin{cases} 0, & \text{if } y = \hat{y} \\ 1, & \text{if } y \neq \hat{y} \end{cases}$$

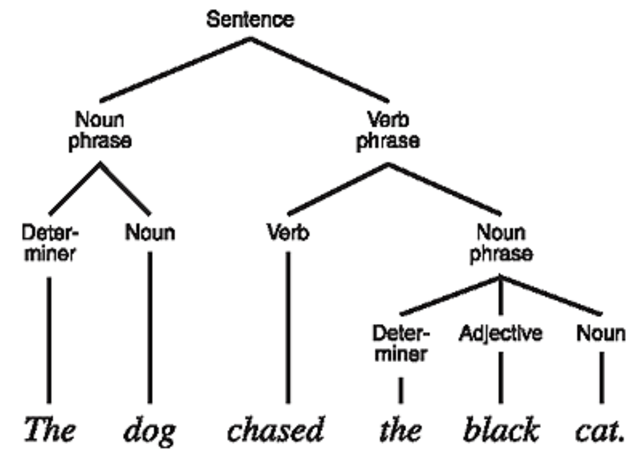
Problem 1: not differentiable wrt $\hat{y}$ (or θ)

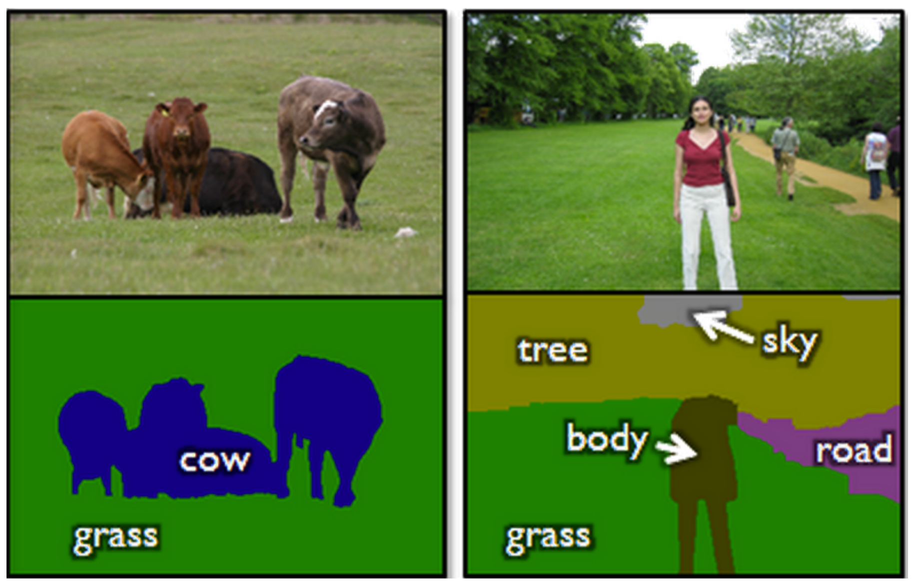
Solution 1: is the data linearly separable? Perceptron (next class) can work

Solution 2: is h(x) a conditional distribution p(y | x)? Maximize that probability (a couple classes)

Structured Classification: Sequence & Structured Prediction







Courtesy Hamed Pirsiavash

Structured Classification Loss Function Example: 0-1 Loss?

$$\ell(y, \hat{y}) = \begin{cases} 0, & \text{if } y = \hat{y} \\ 1, & \text{if } y \neq \hat{y} \end{cases}$$

Problem 1: not differentiable wrt $\hat{y}$ (or θ)

Solution 1: is the data linearly separable? Perceptron (next class) can work

Solution 2: is h(x) a conditional distribution p(y | x)? Use MAP

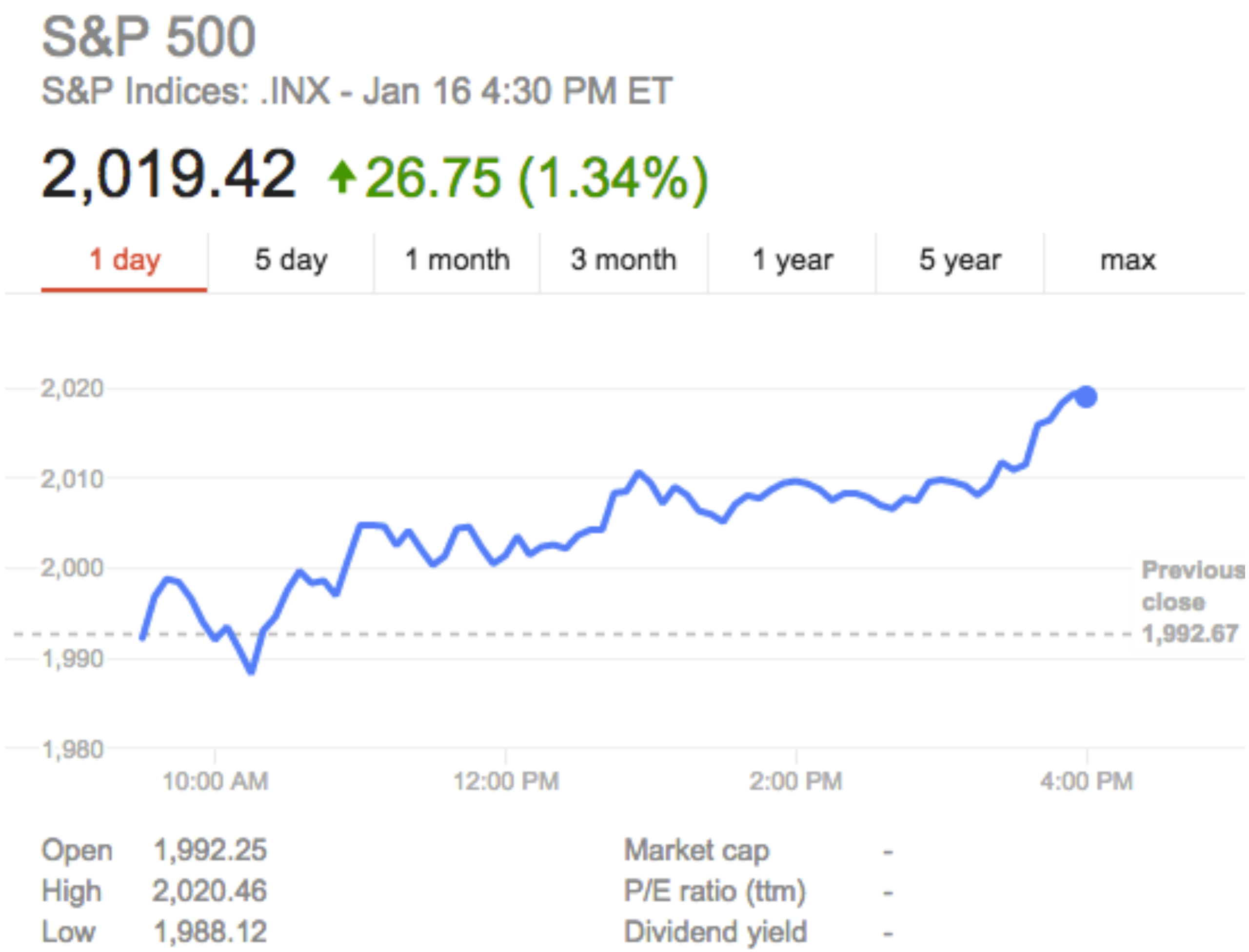
Problem 2: too strict. Structured Prediction involves many individual decisions

Solution 1: Specialize 0-1 to the structured problem at hand

Regression

Like classification, but real-valued

Regression Example: Stock Market Prediction



Courtesy Hamed Pirsiavash

Regression Loss Function Examples

squared loss/MSE
(Mean squared error)

$$\ell(y, \hat{y}) = (\mathrm{y} - \hat{y})^2$$

$\hat{y}$ is a real value → nicely differentiable (generally) ☺

Regression Loss Function Examples

squared loss/MSE
(Mean squared error)

$$\ell(y, \hat{y}) = (\mathrm{y} - \hat{y})^2$$

$\hat{y}$ is a real value → nicely differentiable (generally) ☺

absolute loss

$$\ell(y, \hat{y}) = |y - \hat{y}|$$

Absolute value is *mostly* differentiable

Regression Loss Function Examples

squared loss/MSE
(Mean squared error)

$$\ell(y, \hat{y}) = (\mathrm{y} - \hat{y})^2$$

$\hat{y}$ is a real value → nicely differentiable (generally) ☺

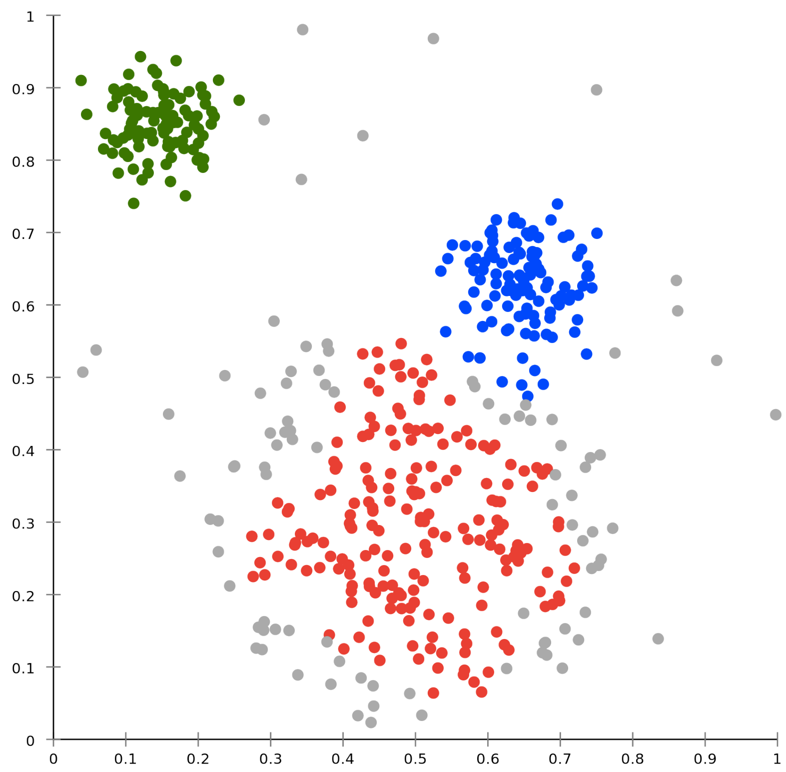
absolute loss

$$\ell(y, \hat{y}) = |y - \hat{y}|$$

Absolute value is *mostly* differentiable

These loss functions prefer different behavior in the predictions (hint: look at the gradient of each)… we'll get back to this

Unsupervised learning: Clustering



We'll return to clustering loss functions later

Courtesy Hamed Pirsiavash

Outline

Review+Extension

Probability

Decision Theory

Loss Functions