CS 412

JAN 16TH – PROBABILITY AND STATISTICS REVIEW

About the course

Midterm

- TBA Next Week, Likely the week before spring break
- 20% of your grade

Final exam (20% of your grade)

Introduction to

Textbooks (all available online as pdf and are optional):

- *The Elements of Statistical Learning,* Hastie, Tibshirani.Friedman (HTF)
- *Hands on Machine learning with Scikit-learn and TensorFlow,* Aurélien Géron
- *(STAT381 review) Mathematical Statistics with Applications,* Wackerly, Mendenhall, Scheaffer

Most lectures will have a corresponding reading—most of them from the HTF book

Why the need for probability?

Assessing “Luck” vs. significant Improvements

Expressing uncertainty in predictions is useful

Continuous random variables

For f(x) to define a valid probability distribution:

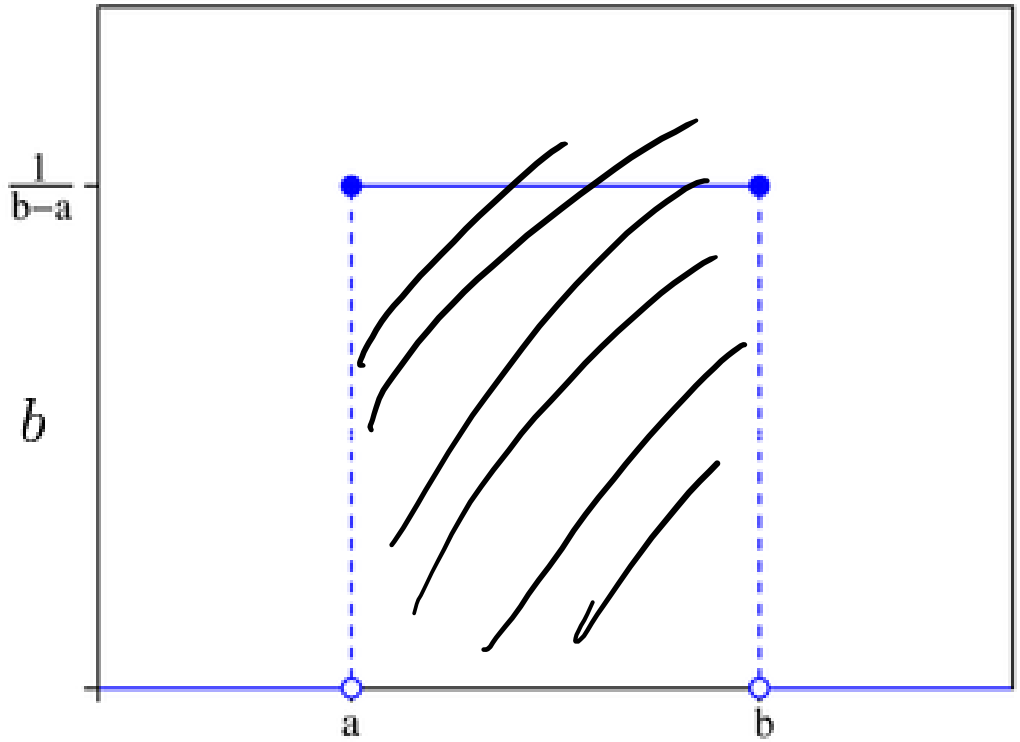
$$f(x) \geq 0$$
$$\int f(x)\, dx = 1$$

Can there exist an x s.t. f(x) > 1.0? Yes

however $\forall_{ab} \int_a^b f(x) \leq 1$

Uniform Distribution

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$

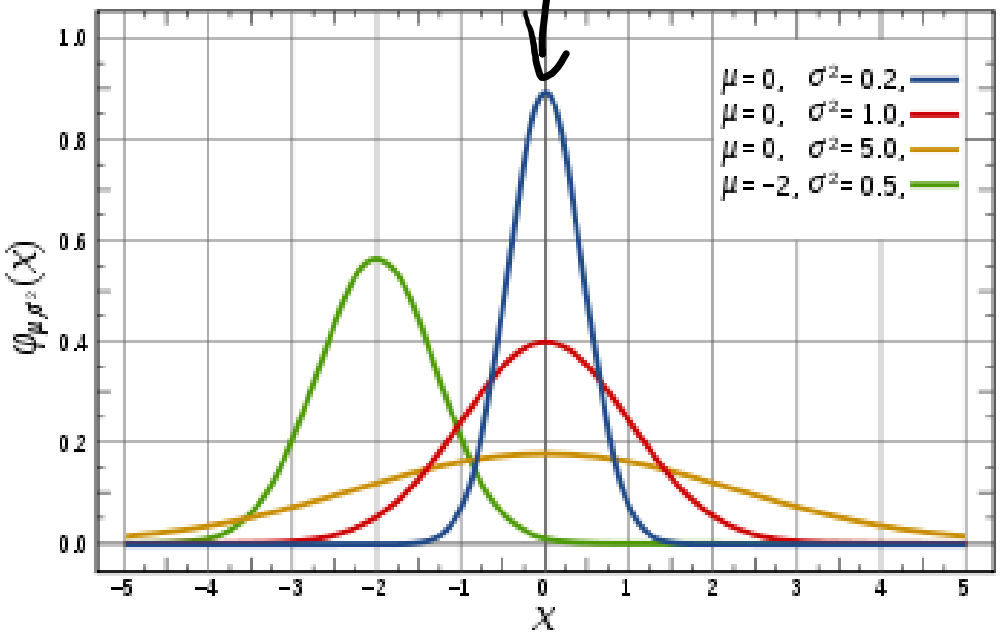


$= 1$

Gaussian Distribution

μ = mean

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

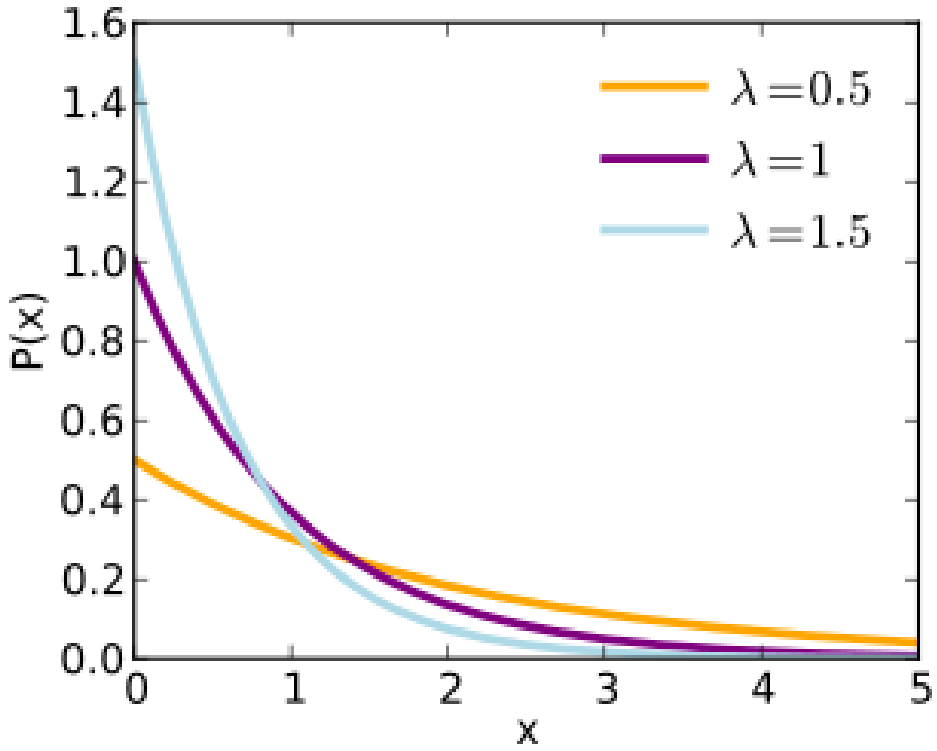


σ = Stdev

σ^2 = variance

Exponential Distribution

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

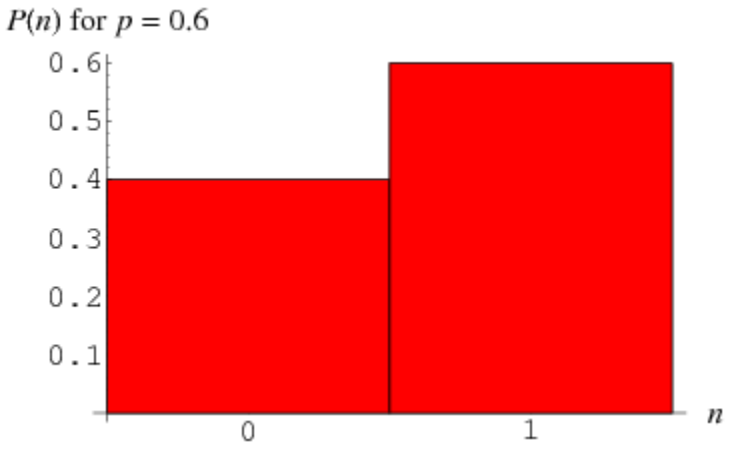


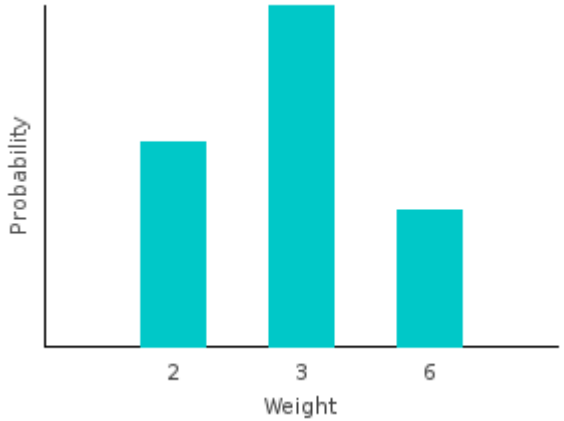
Bernoulli Distribution

$$P(heads) = \theta$$
$$P(tails) = 1 - \theta$$
$$\theta \in [0, 1]$$

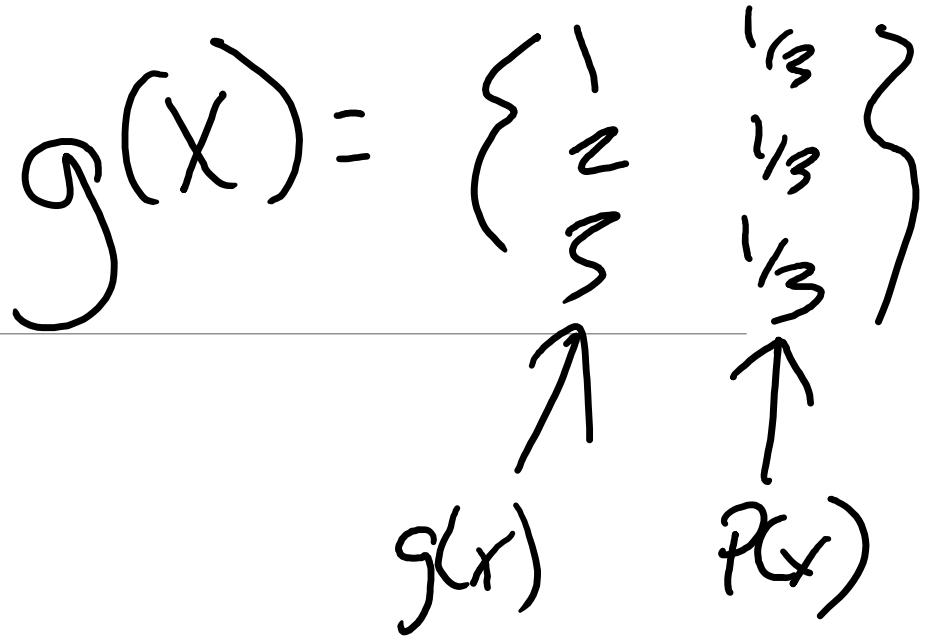
} only two outcomes

Extension: Categorical – given probabilities for multiple events





Expectation



What happens on average?

$$E[g(X)] = \sum P(x)\, g(x)$$

$$(\text{or } \int_x f(x)\, g(x)\, dx)$$

$\theta \cdot 1 + (1-\theta) \cdot 0$

If $X_i \sim \text{Bernoulli}(\theta)$, $E[\sum_{i=1:N} X_i]$? $= \theta \cdot N$

$0 \leq \theta \leq 1$

$\sum P(x) g(x) =$

$1 \cdot 1/3 + 2 \cdot 1/3 + 3 \cdot 1/3$

Linearity properties:

$$E[g_1(X) + g_2(X)] = E[g_1(X)] + E[g_2(X)]$$

$$E[\alpha\, g(X)] = \alpha\, E[g(X)]$$

Mean / Variance / Std. Dev.

Mean: Distribution average, $\mu = E[X]$

Variance: How far distribution is "spread out"

$$\begin{aligned}\sigma^2 &= E[(X - E[X])^2] \\ &= E[(X - \mu)^2] \\ &= E[X^2] - 2\,E[X\,\mu] + \mu^2 \\ &= E[X^2] - \mu^2\end{aligned}$$

Standard Deviation: σ

σ^2

Variance $= \sigma^2$

$E(X - E(X))^2$

Example: Chicago Weather

Random variables: Temp, Sky, Precipitation

Temperature	Sky	Precipitation	Probability
Cold	Clear	No Snow	28.0%
Cold	Clear	Snow	0.0%
Cold	Cloudy	No Snow	8.4%
Cold	Cloudy	Snow	33.6%
Warm	Clear	No Snow	12.0%
Warm	Clear	Snow	0.0%
Warm	Cloudy	No Snow	18.0%
Warm	Cloudy	Snow	0.0%

$\sum_{cold} P(x, y, z)$

What is the probability of being cold?
What is the probability of snow?

Marginal Probabilities

Given a joint probability table, what are the probabilities of a subset of events?

$$P(x) = \Sigma_{y,\,z}\, P(x, y, z)$$

$$P(x, y) = \Sigma_{z}\, P(x, y, z)$$

Conditional Probabilities

Given that one of the random variables has a certain value, what is the distribution of other variables?

$$P(x \mid y, z) = P(x, y, z) / P(y, z)$$

$$P(x, y \mid z) = P(x, y, z) / P(z)$$

Example: Chicago Weather

Temperature	Sky	Precipitation	Probability
Cold	Clear	No Snow	28.0%
Cold	Clear	Snow	0.0%
Cold	Cloudy	No Snow	8.4%
Cold	Cloudy	Snow	33.6%
Warm	Clear	No Snow	12.0%
Warm	Clear	Snow	0.0%
Warm	Cloudy	No Snow	18.0%
Warm	Cloudy	Snow	0.0%

If it is Cold and Cloudy, what is the probability of Snow?
If it is Cold, what is the probability of Snow?

$$P(S \mid C_{\&}, C) = \frac{33.6\%}{33.6+8.4\%}$$

Product Rule

Any joint probability distribution can be written as the product of conditionals:

$$P(x, y, z) = P(x)\, P(y \mid x)\, P(z \mid x, y)$$

$$= P(x)\, [\, P(x, y) / P(x)\,]\; [P(x, y, z) / P(x, y)]$$

Independence

If X is independent of Y ($X \perp Y$) then:

$$P(x|y) = P(x) \text{ and } P(y|x) = P(y)$$

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

equivalent if $P(x)$ and $P(y) \neq 0$

If X is conditionally independent of Y given Z

($X \perp Y | Z$) then:

$$P(x|y,z) = P(x|z) \text{ and } P(y|x,z) = P(y|z)$$

$$P(x,y,z) = P(y,z)\,P(x|z) = P(x,z)\,P(y|z)$$

Example: Chicago Weather

Temperature	Sky	Precipitation	Probability
Cold	Clear	No Snow	28.0%
Cold	Clear	Snow	0.0%
Cold	Cloudy	No Snow	8.4%
Cold	Cloudy	Snow	33.6%
Warm	Clear	No Snow	12.0%
Warm	Clear	Snow	0.0%
Warm	Cloudy	No Snow	18.0%
Warm	Cloudy	Snow	0.0%

Is Temperature independent of Sky?

independent (hopefully)

not conditionally independent given precip

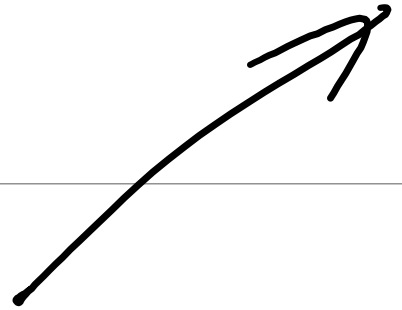
Bayes Theorem

Useful when we know P(y|x), but not P(x|y)

$$P(x|y) = \frac{P(y|x)P(x)}{\sum_{x'} P(y|x')P(x')}$$

$$P(\text{Snow}|\text{Clear}) = \frac{P(\text{Clear}|\text{Snow})\,P(\text{Snow})}{P(\text{Clear}|\text{NoSnow})\,P(\text{noSnow}) + P(\text{Clear}|\text{Snow})\,P(\text{Snow})}$$

Many Trials



Many different trials:

$$T_1, T_2, T_3, \ldots, T_N$$

with joint probability:

$$P(T_1, T_2, T_3, \ldots, T_N)$$

that is often (assumed) independent:

$$P(T_1, T_2, T_3, \ldots, T_N) = P(T_1)\, P(T_2)\, P(T_3) \ldots P(T_N)$$

often assume that they are identically distributed

i.i.d → naive

From Bernoulli to Binomial

$P(0\text{ tails}) = 1/8$
$P(1\text{ tails}) = 3/8$
$P(2\text{ tails}) = 3/8$
$P(3\text{ tails}) = 1/8$

$X_1 \ldots X_n$ i.i.d. Bernoulli random variables θ

What is probability of a specific outcome set?

$$P(x_1 \ldots x_n) = \theta^H (1-\theta)^T$$

for a fair coin

$\theta = 1-\theta = 1/2$

Outcome

$HHT = 1/2^3$

HTH

THH

Outcome Set

1 Tail

What if we care only about the number of heads?

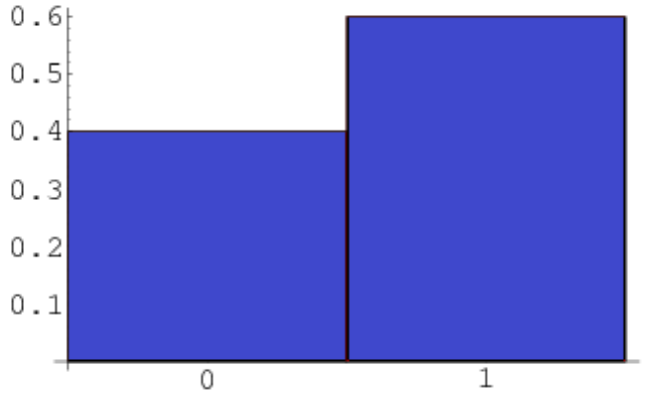
$$P(x_1 + \ldots + x_n = H) = (\text{n choose H})\ \theta^H (1-\theta)^T$$

$$(n,H) = \binom{n}{H} = \frac{n!}{n-H!\cdot H!}$$

Bernoulli Distribution

$$P(x;\theta) = \begin{cases} 1-\theta & \text{for } x = 0 \\ \theta & \text{for } x = 1 \end{cases}$$

$$P(x;\theta) = \theta^{x}\,(1-\theta)^{1-x}$$



Probability questions:

Expected number of "heads" outcomes in 10 flips?

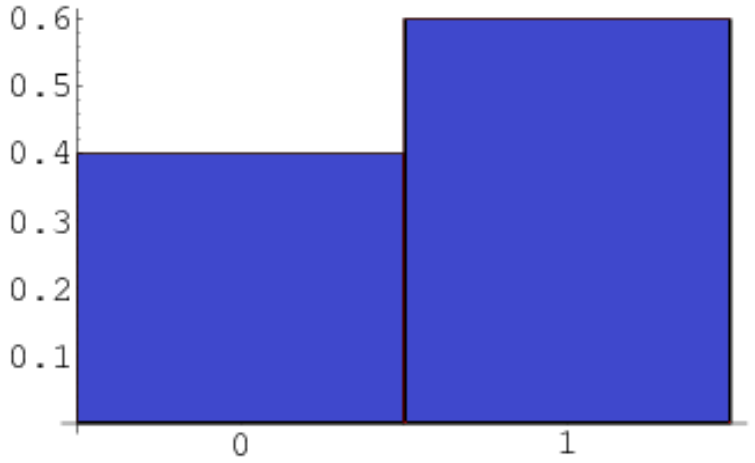
Given 10 flips, what is the probability of at least 8 being "heads?"

$P(\geq 8) = P(8\text{ Heads}) + P(9\text{ Heads}) + P(10\text{ Heads})$

Bernoulli Distribution

$$P(x;\theta) = \begin{cases} 1-\theta & \text{for } x = 0 \\ \theta & \text{for } x = 1 \end{cases}$$

$$P(x;\theta) = \theta^{x}\,(1-\theta)^{1-x}$$



Statistical questions:

Given 3 "heads," 2 "tails," what estimate $\hat{\theta}$?

How far will this estimate be from true θ^*?

Frequentist vs. Bayesian

Frequentist: parameters have fixed values

- Maximum likelihood estimation (MLE) $\hat{}$
- What is the model that is most likely to produce the data (D) given?
- $arg\max_{\theta} P(D \mid \theta)$

Bayesian: parameters are random variables

- Maximum a posteriori (MAP)
- What is the model that is most likely to be the cause of the data?
- $arg\max_{\theta} P(\theta \mid D)$

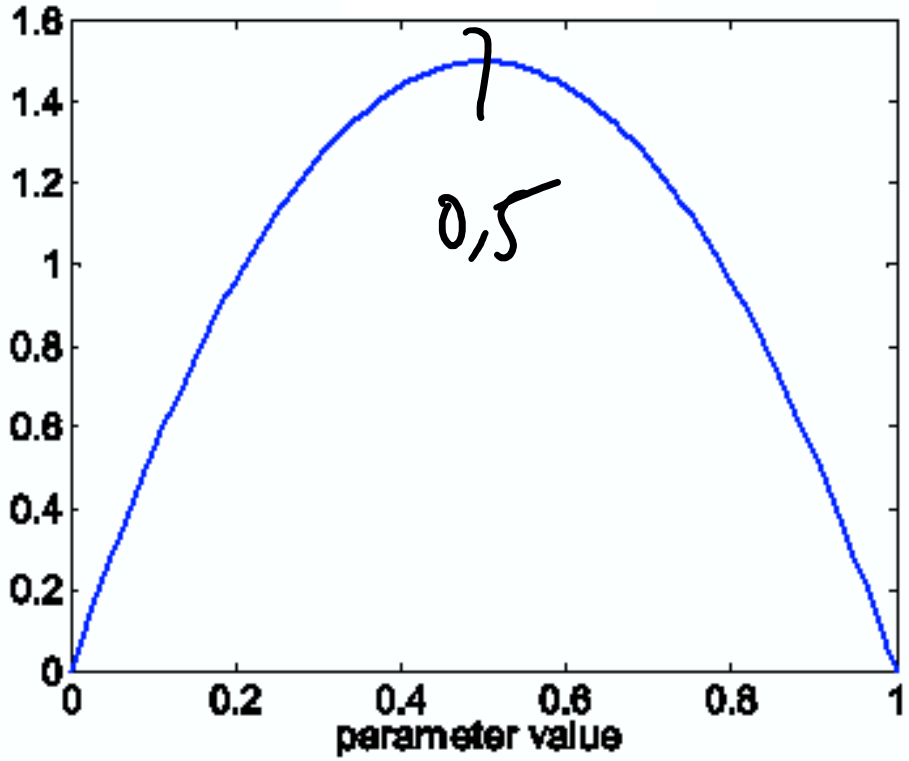
Likelihood Function

Probability of data given the model hypothesis:

$$P(D|\theta)$$

Maximum Likelihood Estimation

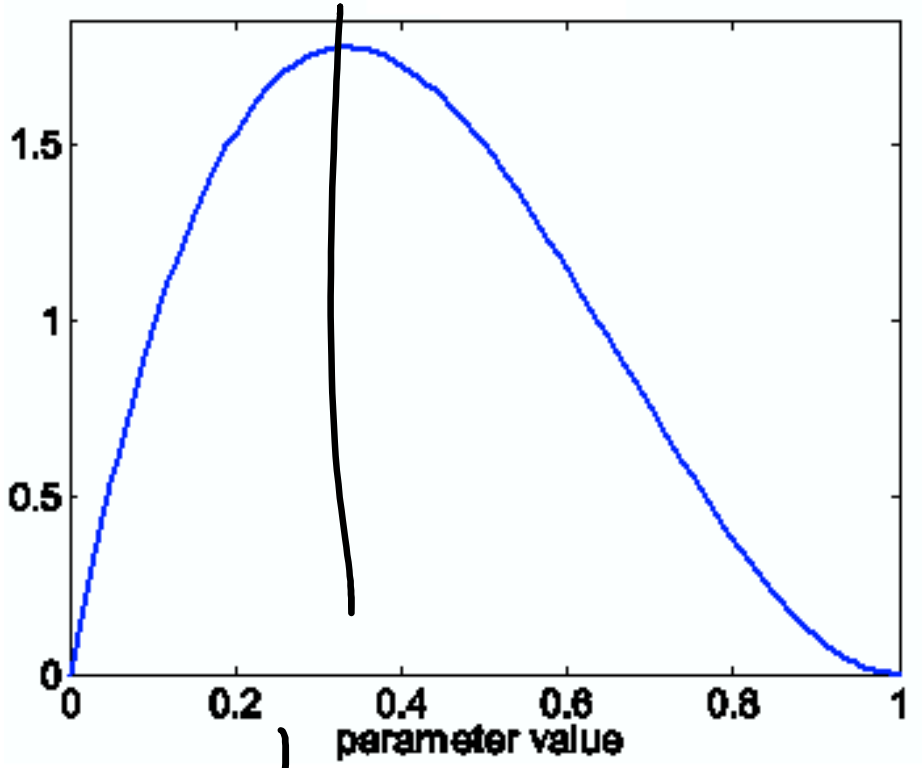
$P(\{1 \text{ head}, 1 \text{ tail}\}; \theta)$



$\hat{\theta} = \text{argmax}_\theta P(D; \theta)$

Maximum Likelihood Estimation

$P(\{1 \text{ head}, 2 \text{ tail}\}; \theta)$



$\hat{\theta} = \text{argmax}_\theta P(D; \theta)$

$E(\hat{\theta} - \theta^*)^2$

Confidence in Estimate?

5 flips: 2 up, 3 down $\theta = 2/5$

50 flips: 20 up, 30 down $\theta = 2/5$

Confidence in Estimate?

5 flips: 2 up, 3 down → $\theta = 2/5$

50 flips: 20 up, 30 down → $\theta = 2/5$

Which is better?

Error Bounds

Can we guarantee our estimate, $\hat{\theta}$, is not far from the true parameter θ^*?

We can never be <u>completely</u> sure…

Probably Approximately Correct (PAC)

via Hoeffding's inequality:

$P(|\hat{\theta} - \theta^*| \geq \epsilon) \leq 2\exp\{-2N\epsilon^2\}$

δ

$E|\hat{\theta} - \theta^*| \leq \epsilon$

ϵ = bound

δ = probability

Error Bounds

$\delta = 0.05$

$\hat{\theta} = 0.5$

To be within ϵ with 95% probability requires N flips:

- $\epsilon = .2$
- $\epsilon = .1$
- $\epsilon = .05$
- $\epsilon = .03$
- $\epsilon = .01$

$\hat{\theta} - \epsilon \leq \theta^* \leq \hat{\theta} + \epsilon$

↳ 95% probability

Error Bounds

$\hat{\theta} = 0.99$

To be within ϵ with 95% probability requires N flips:

$\epsilon = .2$	→	$N = 10$
$\epsilon = .1$	→	$N = 38$
$\epsilon = .05$	→	$N = 149$
$\epsilon = .03$	→	$N = 414$
$\epsilon = .01$	→	$N = 931$

$\hat{\theta} \leq 0.99 + 0.2$

Summary

Probability important for:

- Estimating prediction uncertainty
- Analyzing random samples

Many important concepts:

- Random variables
- Marginalization
- Approximation bounds PAC

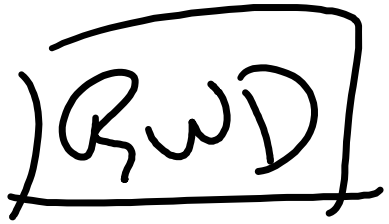
Modeling

Probability important for:

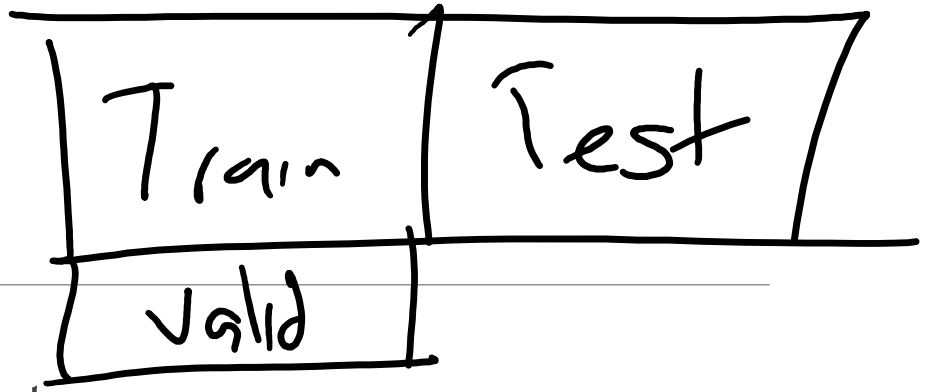
- Estimating prediction uncertainty
- Analyzing random samples

Many important concepts:

- Random variables
- Marginalization
- Approximation bounds



Data Separation



Before your model construction begins,
you need to separate your data into at least two sets

- Training
- Testing
- *Validation (which we'll discuss later)*

You should not conduct any analysis of the data before separating it

- This is called *data snooping* and can result in an inaccurate estimate of our final reported error

The only way we can estimate the final reported error of our model
is by judging on the test set

Data Separation

The training set is used to train the data and is usually a smaller section of the data as a whole (much less than half)

By not analyzing or training on a large set of the data, what is the advantage?

Data Separation

The training set is used to train the data and is usually a smaller section of the data as a whole (much less than half)

By not analyzing or training on a large set of the data, what is the advantage?

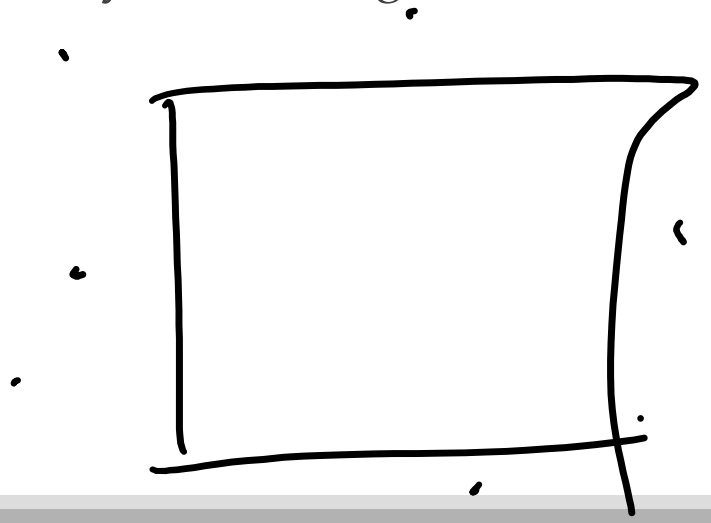
When we've created our "choice" model, we can run it on the test data *only once* and get an estimate of the error without having to induce a penalty

Approach

A bank creates an investment ML algorithm that forcasts whether a certain bond value will go up or down. As input, it receives a group of financial features that have all been standardized for mean and variance. The bank separated the data into two groups, training and testing. After building the model on the training data, the model trained on the testing data can correctly determine whether the bond will go up or down with 52% accuracy.

only one

When this model was applied to the actual market, it consistently was wrong less than 50% of the time. What went wrong?

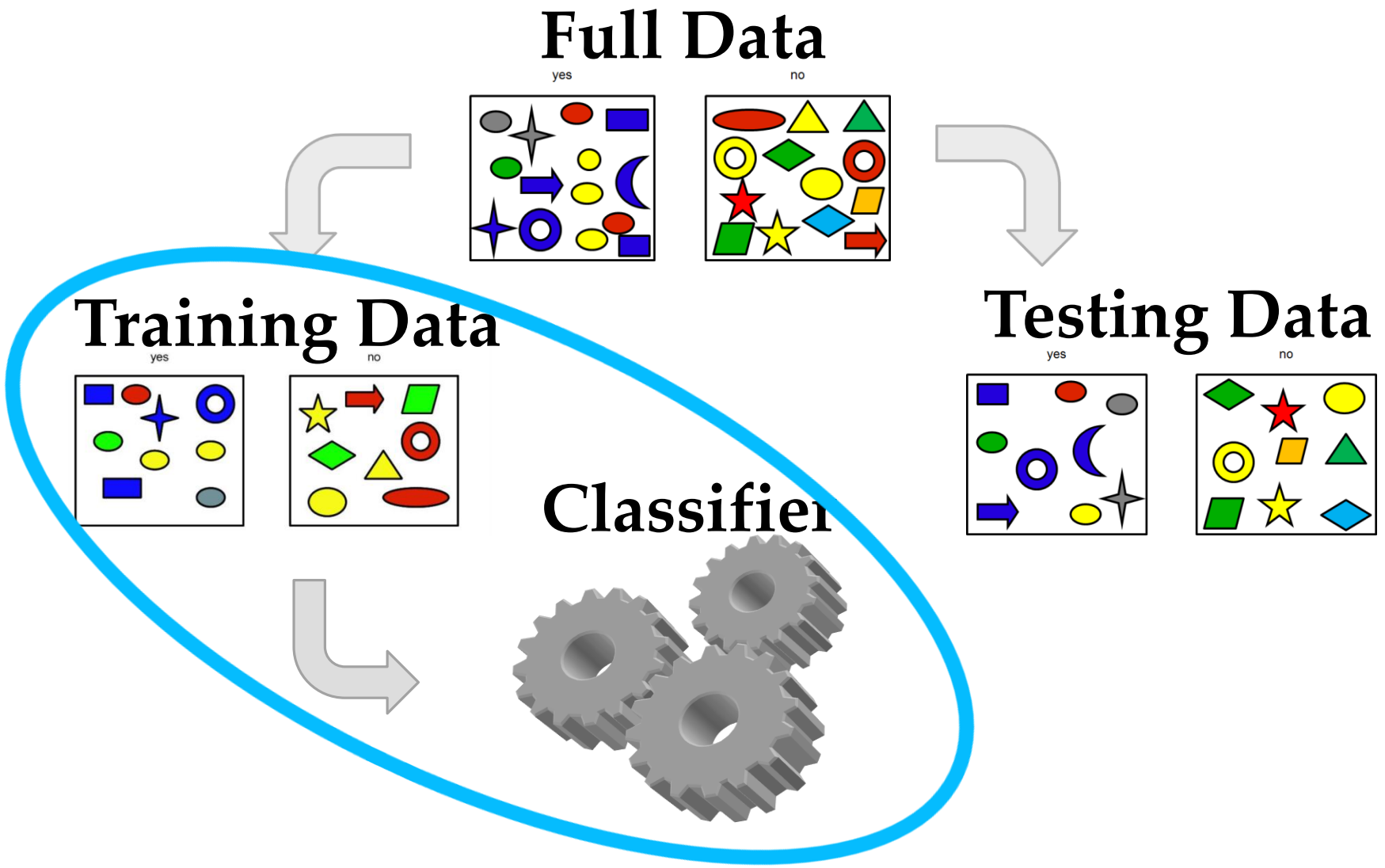


Approach

A bank creates an investment ML algorithm that forcasts whether a certain bond value will go up or down. As input, it receives a group of financial features that have all been standardized for mean and variance. The bank separated the data into two groups, training and testing. After building the model on the training data, the model trained on the testing data can correctly determine whether the bond will go up or down with 52% accuracy.

When this model was applied to the actual market, it consistently was wrong less than 50% of the time. What went wrong?

The bank normalized both the training and testing sets, meaning the training set was actually impacted by the values in the test set.


Full Data
yes
no
Training Data
yes
no
Classifier
Testing Data
yes
no