EE5907 Pattern Recognition
EE5027 Statistical Pattern Recognition

BT Thomas Yeo

ECE, CSC, TMR, N.1, HMS

Topics

- Thomas Yeo (EE5907 Part I + EE5027)
  - Contact: thomas.yeo@nus.edu.sg
  - Bayesian decision theory
  - Parameter estimation & supervised learning
  - Non-parametric techniques
- Robby Tan, Mike Shou, Wang Xinchao (EE5907 Part II + EE5026)
  - Unsupervised / supervised dimensionality reduction
  - Clustering and Applications
  - Deep learning

(Rough) Schedule

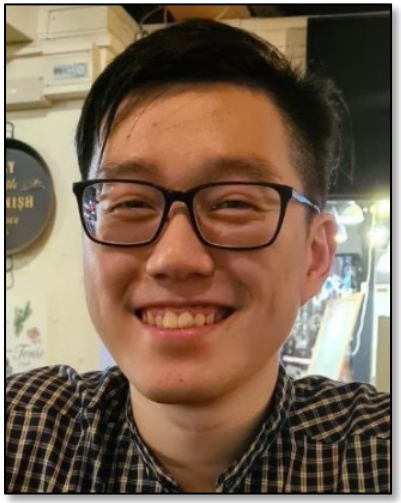
Week 1	Introduction + Probability Review
Week 2	Parameter Estimation (ML, MAP, Bayesian) Generative & Discriminative models Conjugate Distributions
Week 3	Univariate Gaussian + Parameter Estimation (ML, MAP, Bayesian) Naive Bayes
Week 4	Supervised learning (logistic regression)
Week 5	Non-parametric techniques (KNN, Parzen's window)
Week 6	Bayesian Statistics / Decision Theory

Hours

- Lecture: 6pm – 9pm (Thurs) @ Zoom

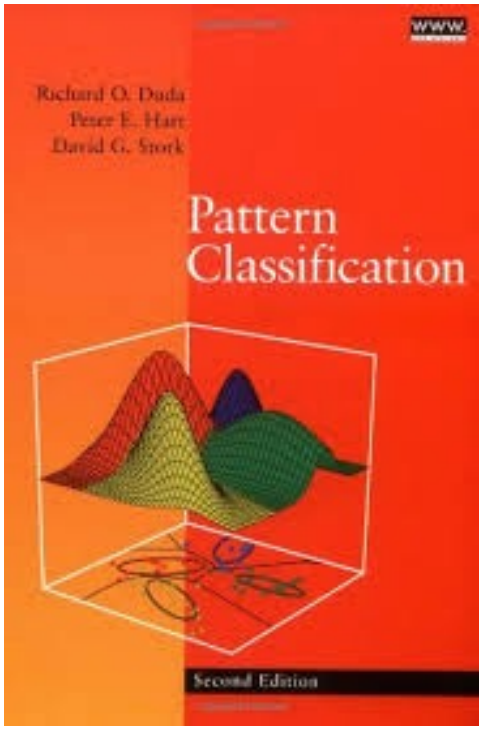
- Thomas' office hours:
  - After lecture @ Zoom
  - For second half of semester, by appointment only

TA: EE5027 + first half of EE5907

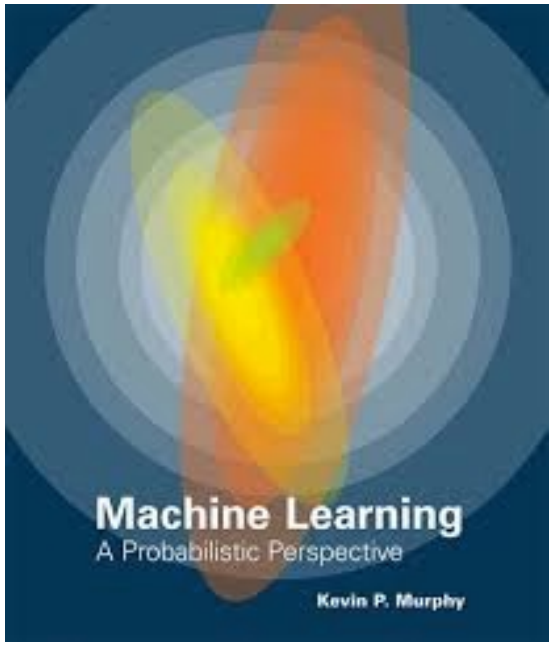


- Office hours: 6pm – 7pm (Monday) @ Zoom or by appointment
  - Zoom meeting ID: 851 7898 3125
  - Password: 001
- Leon Ooi (e0471099@u.nus.edu)
- Yan Xiaoxuan (e0012668@u.nus.edu)
- For students appealing to join the class, please email Leon or Xiaoxuan your email ID (the one starting with “e”) and they will add you as guests on LumiNUS. Remember to cc me.

Textbook & References



Pattern Classification by Duda, Hart and Stork, John Wiley, 2001.



Machine Learning: A Probabilistic Perspective, Kevin Murphy, 2012

Computer Vision: Models, learning and inference, Simon Prince, 2012
Free: www.computervisionmodels.com

Pre-requisites

- Pre-requisites
  - Probability, statistics, and linear algebra (vector spaces and matrix theory) as taught in typical undergraduate courses
  - Probability and statistics are especially important for first half of EE5907 + EE5027
  - Programming in MATLAB/Python, or standard languages such as C, C++, Java, etc
- You will find this class very difficult if you don't have the pre-requisites
- If this is your first graduate class, you will also find it to be substantially more difficult than your undergrad courses

State-of-the-Art

- EE5907/EE5027/EE5026 do NOT cover state-of-the-art. Teach concepts needed to understand state-of-the-art

- For state-of-the-art:
  - Conference papers: CVPR, ICCV, ECCV, ICML, ICLR, NeurIPS, etc
  - Journal articles: IJCV, T-IP, JMLR, T-PAMI, etc
  - EE6733: Advanced Topics on Vision and Machine Learning

- EE5934/EE6934: Deep Learning

LuminNUS

- Lecture slides
  - Condensed version in PDF format (you can search for terms in this format)
  - Long PPT version with detailed notes
- Lecture video will be available on LuminNUS
- Assignments & solutions
  - Assignment starts this very week!
  - Assignments not graded, but you should do them!
  - Will go through solutions of some problems in class
  - Solutions posted one week after assignment

Assessments

- EE5907
  - 2 CAs (2 x 20%)
    - Individual projects (absolutely no copying permitted)
  - Final Exam (60%)
    - More information to come
- EE5027
  - 1 CA (40%) same as EE5907 CA1
  - Final exam (60%)
    - More information to come

CA1 (EE5907 + EE5027)

- CA1 handed out on week 3
- Due Monday after recess week
- Worth 20% of your grades for EE5907 & 40% of your grades for EE5027
- Four sections following lectures 3 to 5.
  - Lectures teach theoretical concepts
  - Ungraded homework helps you understand theoretical concepts
  - CA1 lets you implement these theoretical concepts
- Please do NOT wait till last minute
  - According to students from previous years: average = 34 hours (min = 5 hours, max = 120 hours)

Introduction

What is Artificial Intelligence (AI)?

- AI = intelligence displayed by machines
- Unsolved since 1950s
- Recent impressive results by universities (e.g., UoT, NYU, etc) and companies (e.g., Deepmind, Google)
- Resurgence made possible by machine learning

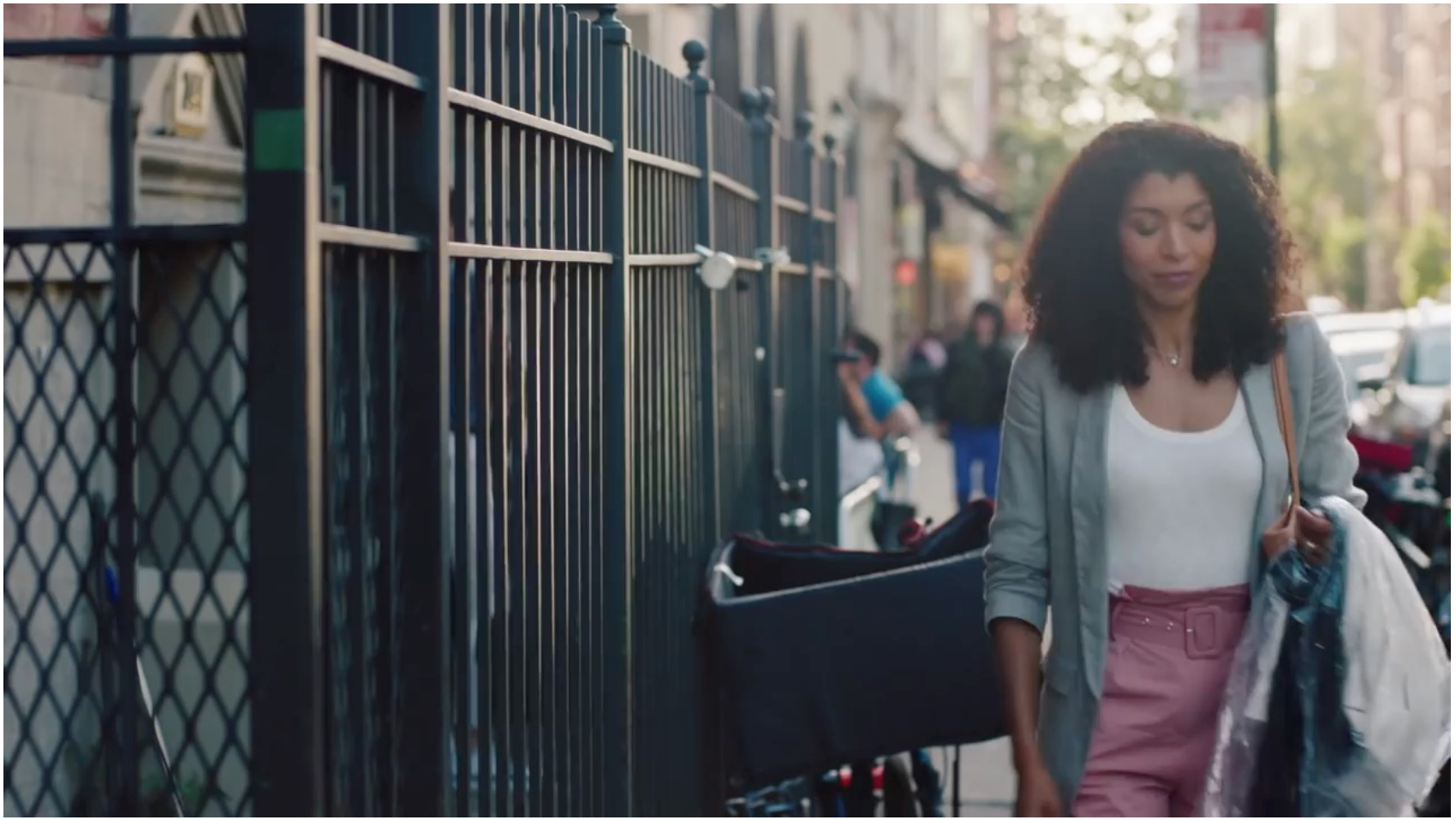
What is Pattern Recognition / Machine Learning?

- Will use both terms interchangeably
  - Pattern Recognition: automatically detect patterns from data
  - Machine Learning: Learn from data without being explicitly programmed
- Applications
  - Email SPAM detection
  - Speech recognition (e.g., SIRI)
  - Handwritten digit recognition (e.g., sorting snail mail)
  - Google search
  - Driverless cars
  - Banking fraud detection

What is Pattern Recognition / Machine Learning?

- Will use both terms interchangeably
  - Pattern Recognition: automatically detect patterns from data
  - Machine Learning: Learn from data without being explicitly programmed
- Applications
  - Email SPAM detection
  - Speech recognition (e.g., SIRI)
  - Handwritten digit recognition (e.g., sorting snail mail)
  - Google search
  - Driverless cars
  - Banking fraud detection
  - Detection and diagnosis of diseases
  - Biometric (e.g., fingerprint security)
  - DNA sequence identification (e.g., Counsyl)
  - Manufacturing (e.g., machine vision)

Google Assistant



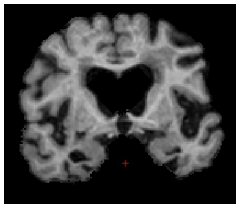
Types of Machine Learning

Supervised Learning

- Given input-output pairs $D = \{x_i, y_i\}_{i=1:N}$, learn mapping $y = f(x)$
  - D = training set
  - x (images, emails, molecular shapes, etc) typically represented as p-dimensional vector (called features)
  - y also called output or target variable
    - y discrete => problem known as classification
    - y continuous => problem known as regression

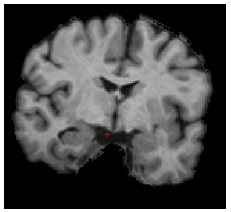
Training: Learn relationship between inputs (MRI) & target labels (AD or healthy)

Alzheimer's Disease (AD)

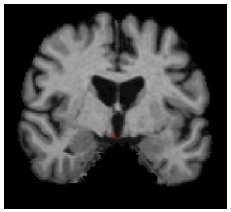


...

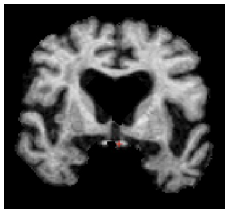
Healthy



...



Testing: Given new MRI, Predict AD or healthy



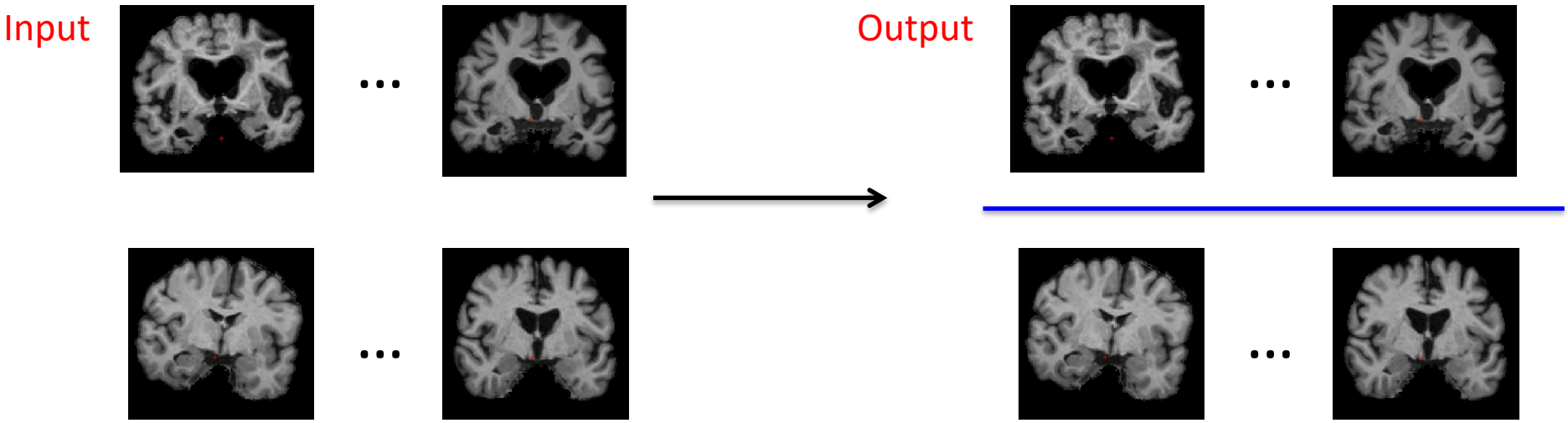
Healthy or AD?

Healthy or AD?

x = MRI images, y = AD or healthy

Unsupervised Learning

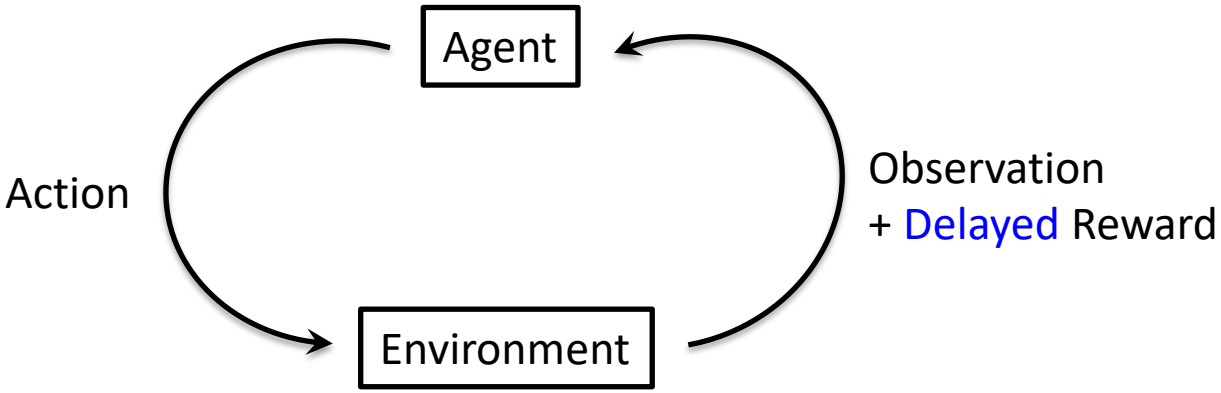
- Given $D = \{x_i\}_{i=1:N}$, discover "interesting structure" in data
  - x (images, emails, molecular shapes, etc) typically represented as p-dimensional vectors
  - "cheap" since labels (i.e., y's) typically expensive to get
  - More open-ended, no groundtruth – validation more difficult



x = MRI images

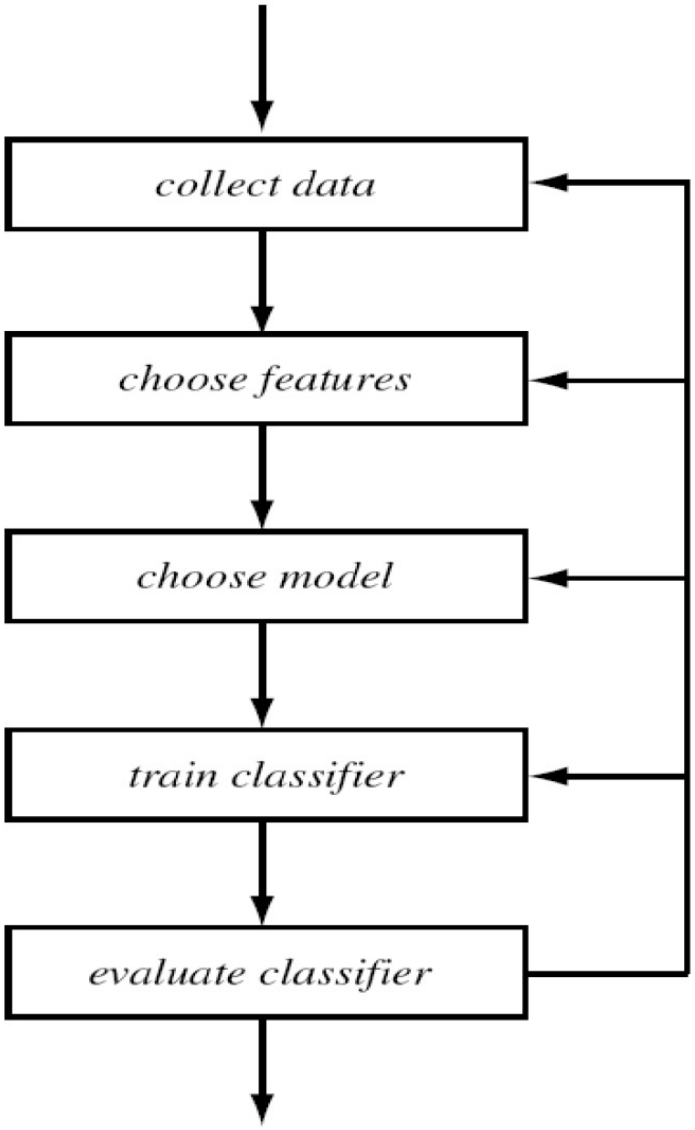
Types of Machine Learning

- Summary
  - Supervised learning: Great if we know what we want to predict
  - Unsupervised learning: Great if we want to discover something new
- Other learning not covered in this class
  - Semi-supervised: some data has target labels, some don't
  - Reinforcement learning: learn actions by trial & error

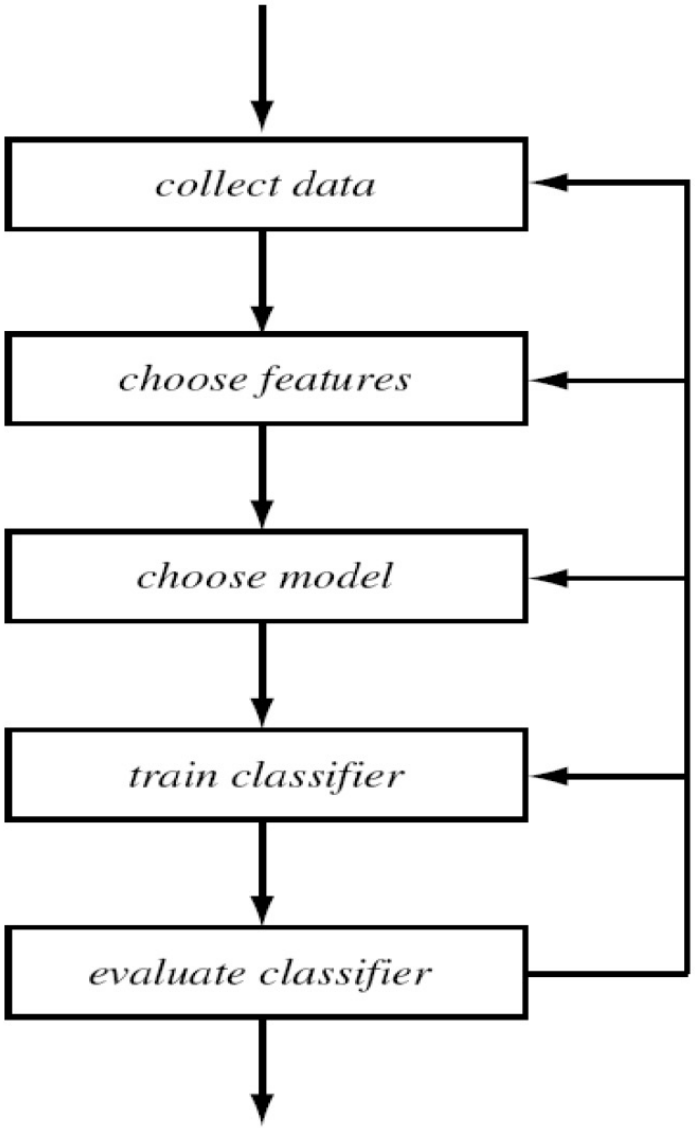


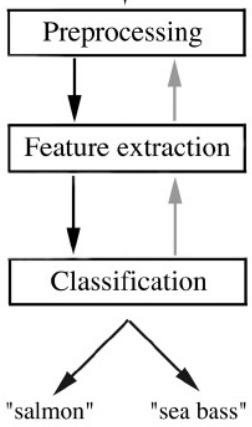
Some Background Knowledge in Machine Learning

Design Cycle For Supervised Learning



Design Cycle Example: Salmon vs Seabass

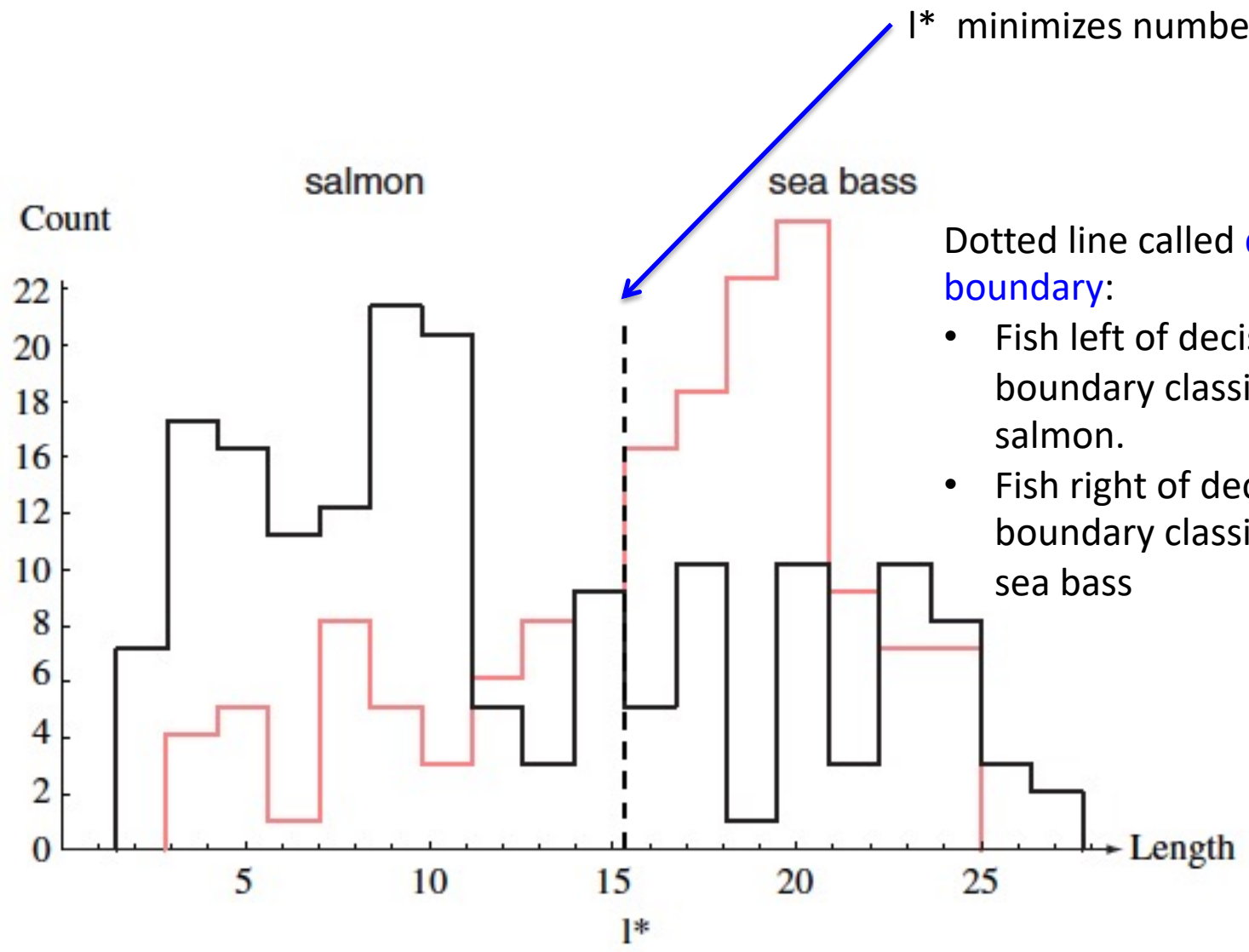




Design Cycle Example: Salmon vs Seabass

- Data collection
  - Set up convey belt and cameras, capture images
- Preprocessing
  - Image enhancement, remove background, separate occluding fishes, extract single fish (segmentation)
- Divide data into training and test sets
  - Feature extraction/engineering, e.g., measure certain features of fish to be used for classifier
  - Train classifier on training set and evaluate on test set
- Re-visit previous steps if performance unsatisfactory

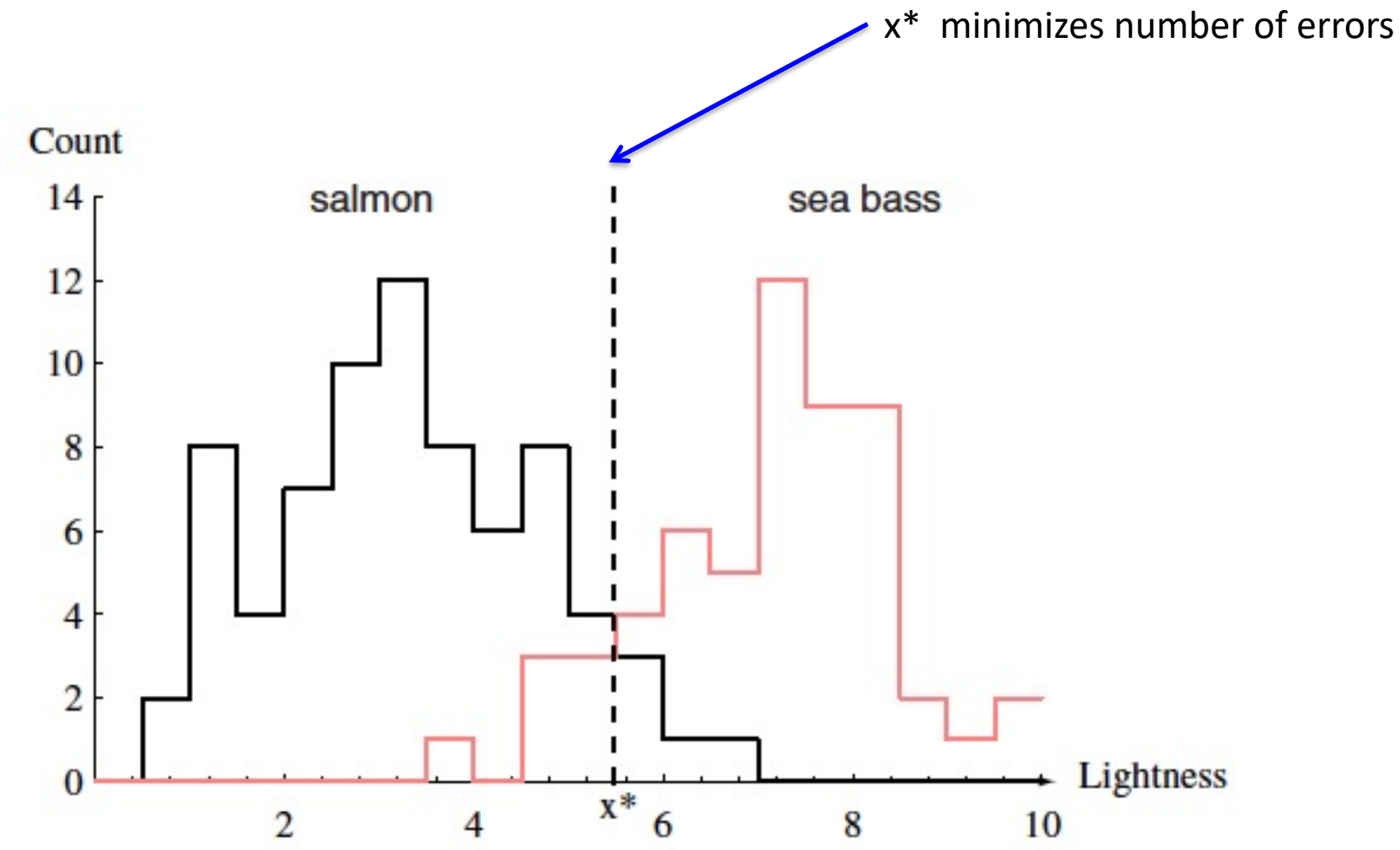
Feature Engineering is Important!



Dotted line called decision boundary:

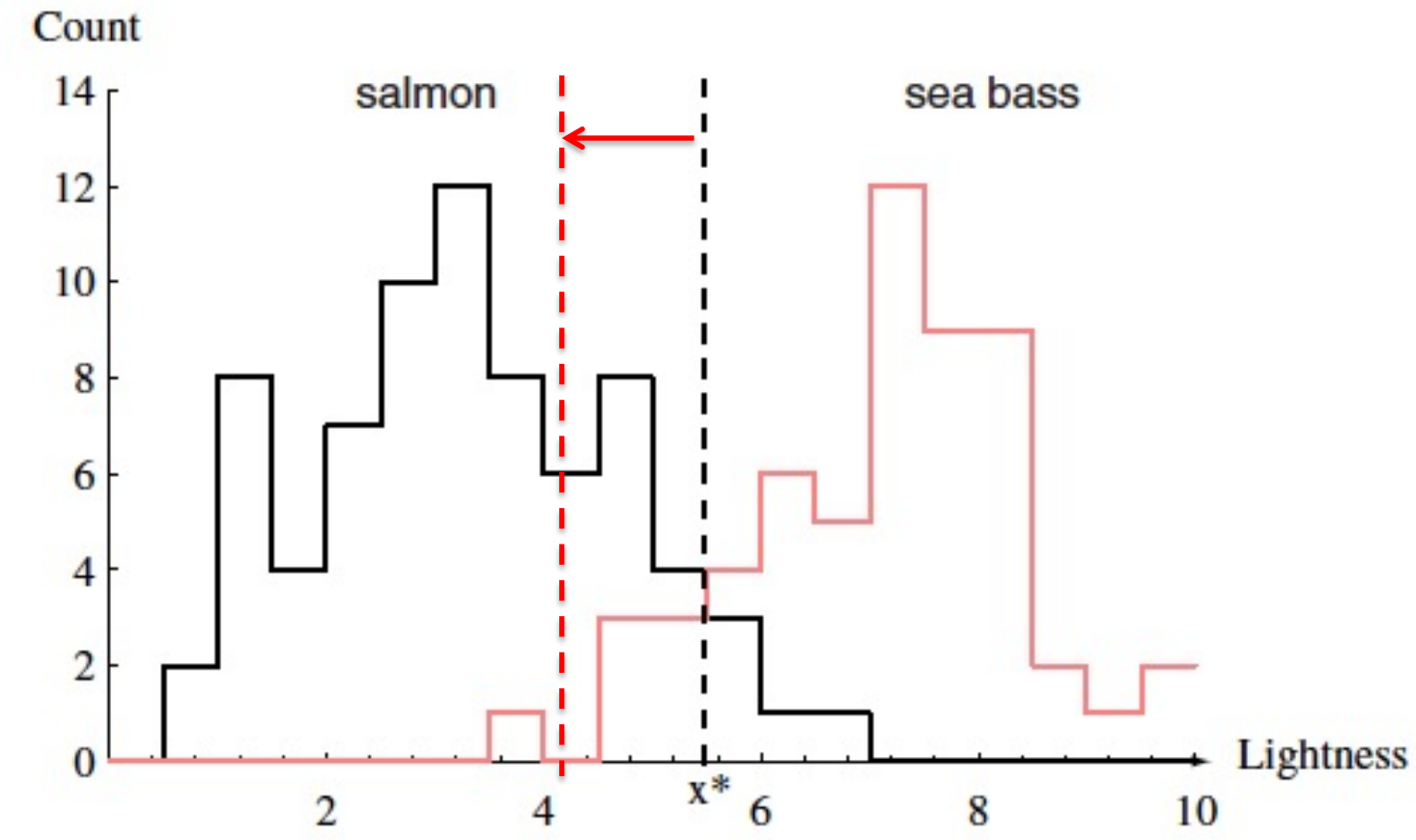
- Fish left of decision boundary classified as salmon.
- Fish right of decision boundary classified as sea bass

Feature Engineering is Important!



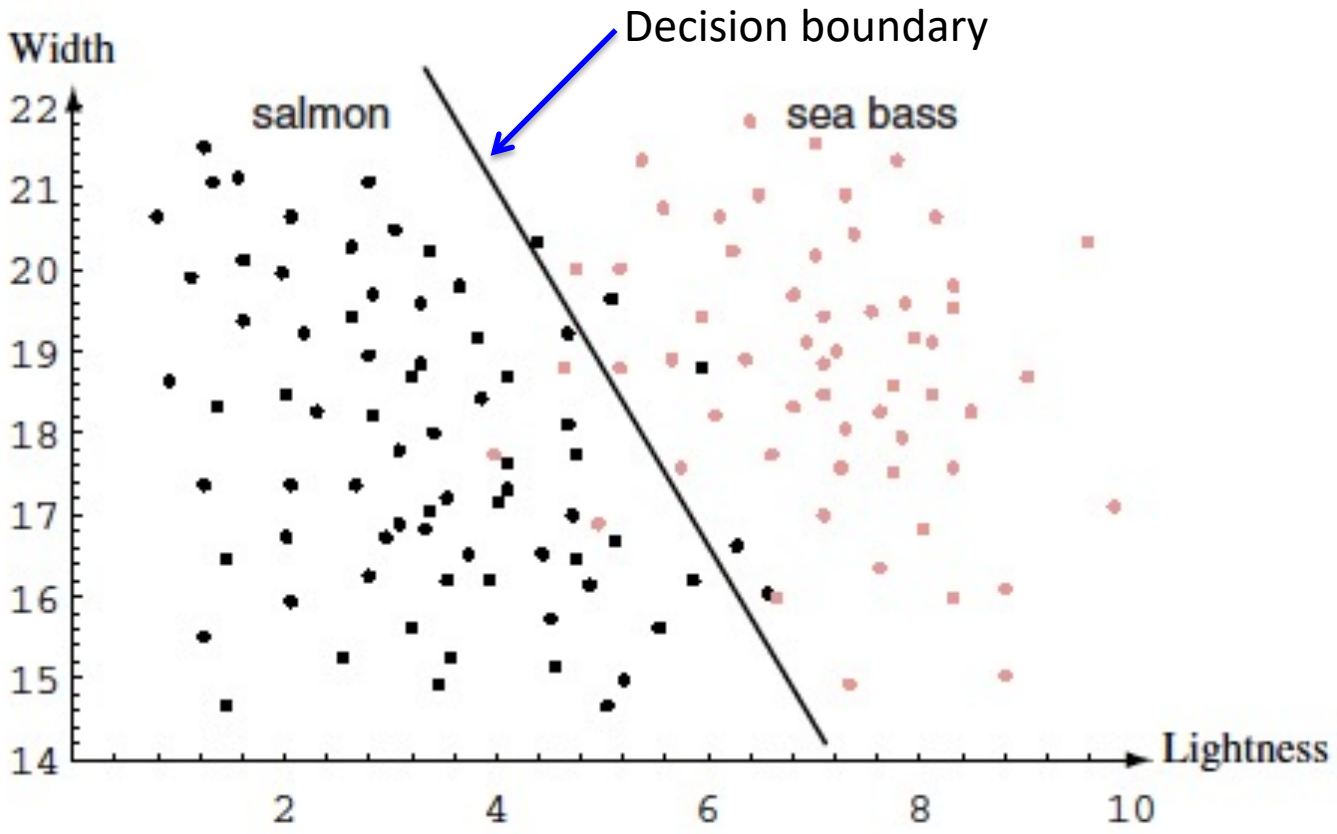
Decision Theory: Not All Errors are Equal

- Example: customers ok with finding salmon in cans marked as sea bass, but very upset to find sea bass in cans marked as salmon
  - Shift decision boundary to minimize angry customers



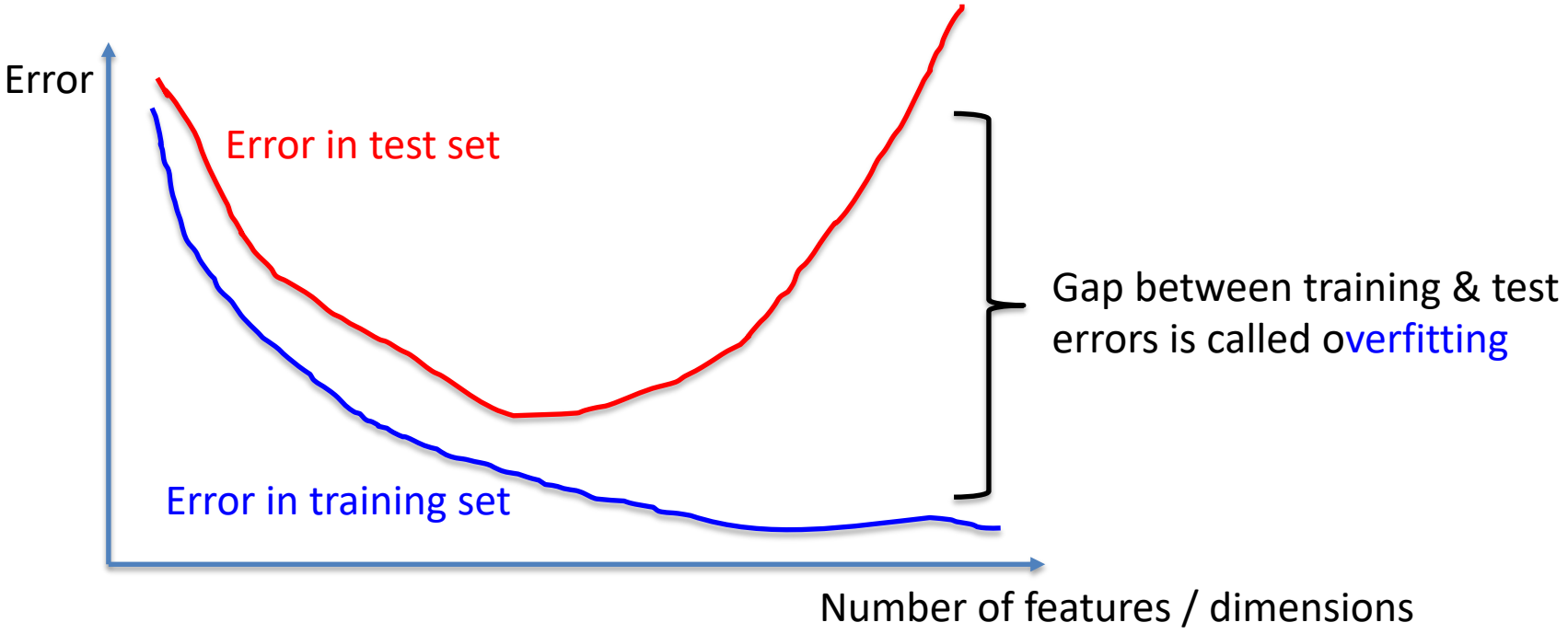
Curse of Dimensionality

- More features might improve performance, but some features (e.g., length) might be useless
- Curse of dimensionality: too many features can lead to worse performance



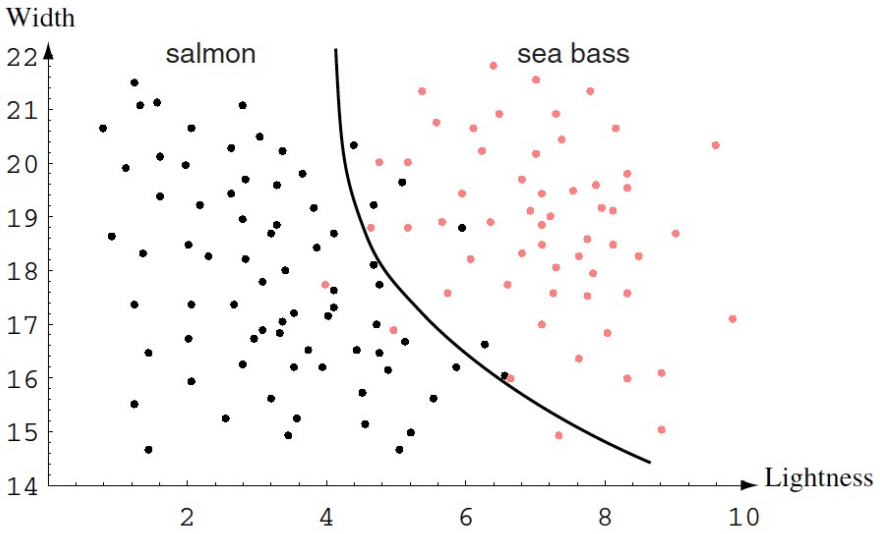
Curse of Dimensionality

- For fixed amount of data samples N (e.g., number of fish photos), as we increase number of features
  - Training error might keep decreasing
  - Test error might decrease and then increase

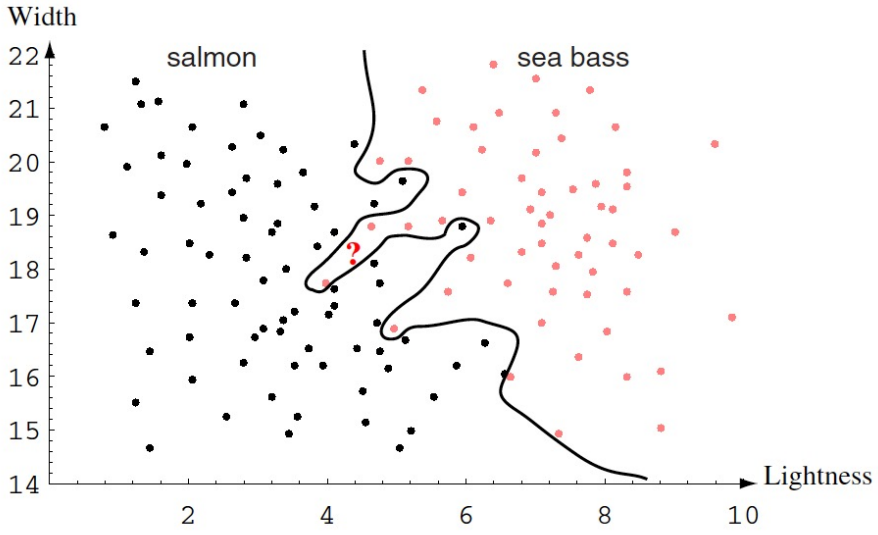


Model Complexity

- Instead of increasing the number of features, can also use more complex decision boundaries (models)
- This can potentially reduce errors, but might also increase errors
- Model complexity roughly related to # model parameters (e.g., quadratic decision boundary requires more parameters to specify than linear decision boundary)



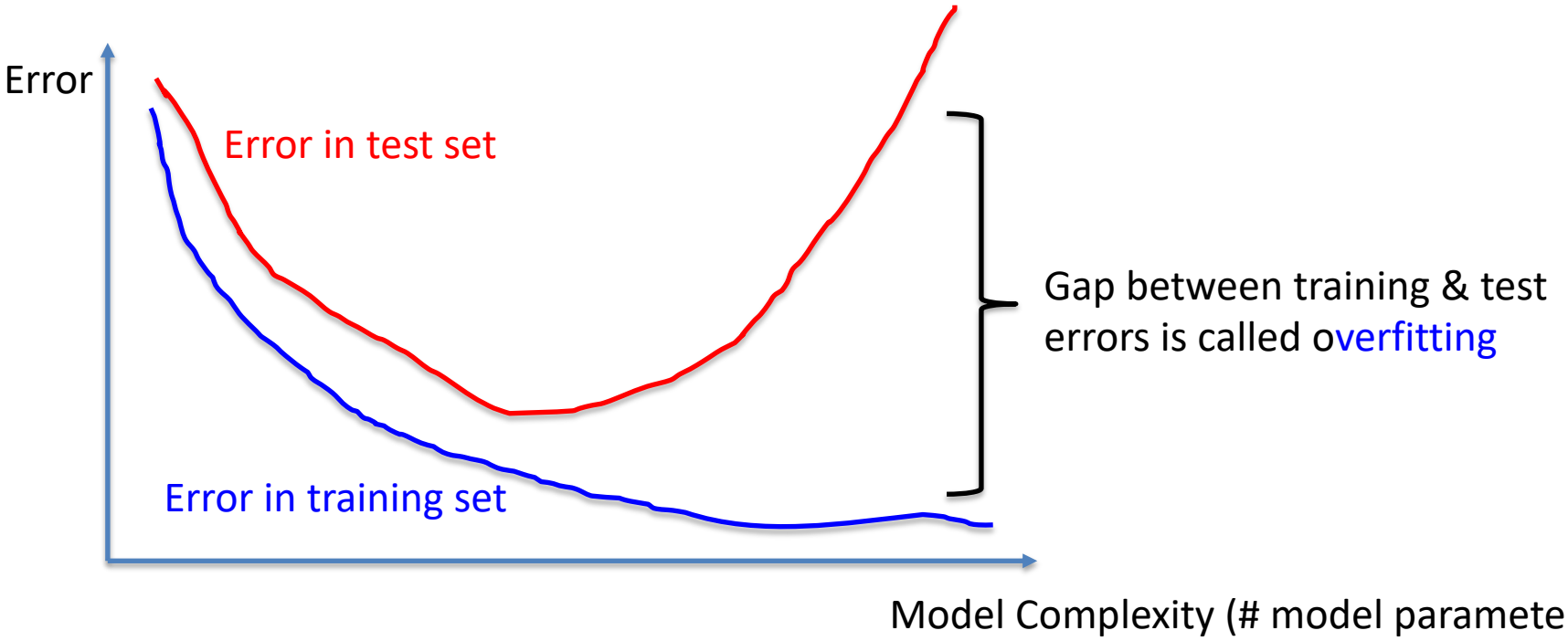
Shifting from linear straight line to quadratic curve reduces errors (in this example)



"?" likely to be salmon, but classified as sea bass

Model Complexity

- For fixed amount of data samples N (e.g., number of fish photos), as we increase number of model parameters
  - Training error might keep decreasing
  - Test error might decrease and then increase



Training, Validation, Test Sets

- To really test the quality of our algorithm, need to evaluate generalization error on test set
- However, splitting data into training-test set not enough
  - Imagine we train an algorithm on training set and error is terrible in test set
  - We then train new algorithm on training set and error is now better in test set
  - But we will have used the test set twice. If you repeat this many times, you will overfit to the test set
- Best practice: split data into training, validation and test sets
  - Train on training set & evaluate error on validation set
  - Repeat as many times as we like
  - Once satisfied with results, then apply to test set to get realistic error quantification

K-fold Cross-Validation

- Sometimes not enough data => K-fold cross validation
  - Divide data into K sets (called folds)
  - Repeat for i = 1 to K
    - Take the i-th fold and call it the test fold
    - Train on remaining folds and test on i-th test fold
  - If K is equal to N (# data points) => leave-one-out cross-validation
- Same problem as before
  - If one algorithm gives poor K-fold cross-validation errors, and we repeat with a different algorithm, we will eventually overfit
  - Solution: Inner-loop (Nested) Cross-Validation
  - Repeat for i = 1 to K
    - Take the i-th fold and call it the test fold
    - Performing K'-fold cross-validation on remaining folds with different algorithms
    - Apply best algorithm to i-th test fold
- K-fold cross-validation more data efficient than training-validation-test, but more complicated and less "clean"

Computational Complexity is Important

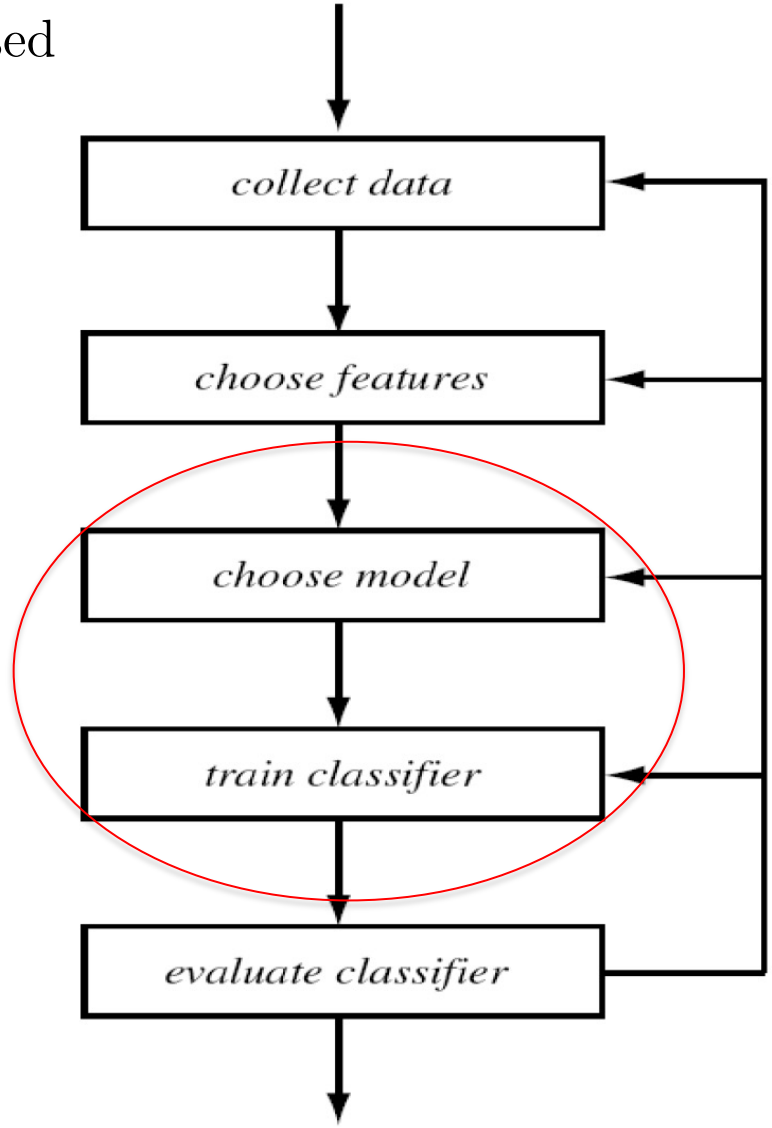
- Netflix million dollar contest to predict movies people enjoy watching
  - Ended up using simpler algorithm with lower accuracy

No Free Lunch Theorem

- "All models are wrong, but some models are useful" – George Box
- No free lunch theorem (Wolpert 1996) – there is no best model that works well for ALL problems
- Therefore need to develop different models/algorithms to cover different kind of data and problems

What will be covered?

- Focus on classification/regression/unsupervised learning
  - Assume features already extracted
  - Strong bias on probabilistic approaches



Why Not Just Teach Deep Learning?

- In certain settings, deep learning do not beat classical machine learning, e.g., logistic regression
  - These classical approaches can be significantly faster and easier to implement

	Hospital A
Inpatient Mortality, AUROC[1](95% CI)	
Deep learning 24 hours after admission	**0.95**(0.94-0.96)
Full feature enhanced baseline at 24 hours after admission	0.93 (0.92-0.95)

Deep learning

Logistic Regression

Google Research, Scalable and accurate deep learning with electronic health records, NPJ Digital Medicine, 2018

- Machine learning is cyclical
  - Our goal not to teach you only the popular stuff, but foundational knowledge useful regardless of what is popular because that will change

Interim Summary

- Different types of learning: supervised (classification vs regression), unsupervised, semi-supervised, reinforcement learning
- Iterative nature: collect data, choose features, choose models, train, test, rinse and repeat
- Feature engineering (i.e., representation) is important
- Decision theory: not all errors are equal
- Curse of dimensionality, model complexity, overfitting
- Training-validation-testing
- K-fold cross validation
- Computational Complexity
- No free lunch

Optional Reading I

- These lecture notes based mostly D&H Chapter 1

- KM Chapter 1 provides different introduction
  - More real world examples
  - Beware of typos

Probability Review

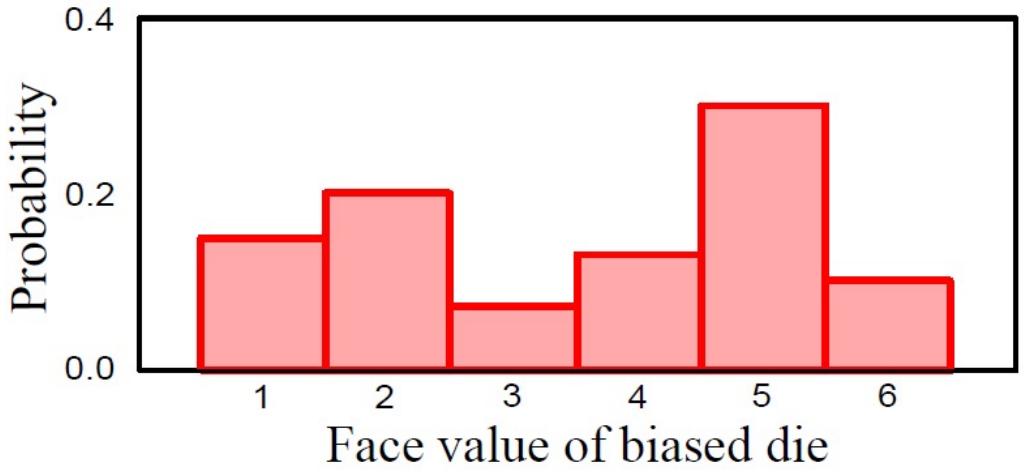
What is a random variable?

- A random variable x is a quantity that is uncertain
- May be result of experiment (e.g., flipping a coin) or real world measurement (e.g., measuring temperature)
- If observe x multiple times, we get different values
- Some values occur more than others; this information captured by probability distribution p(x)
- If x is discrete, then “p” is “probability mass function” (or pmf). If x is continuous, then “p” is “probability distribution function” (or pdf).

Adapted from S. Prince

Discrete Random Variable x

- Take on discrete values

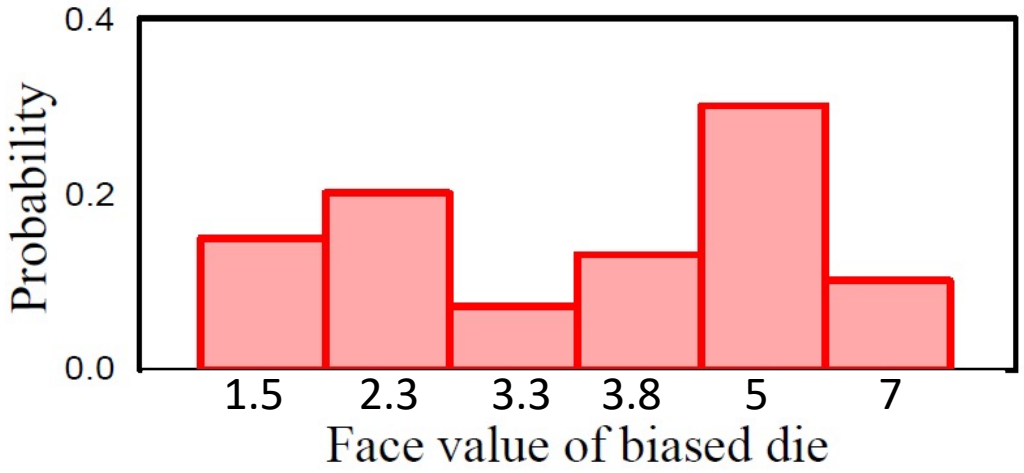


$$\sum_x p(x) = 1$$

Adapted from S. Prince

Discrete Random Variable x

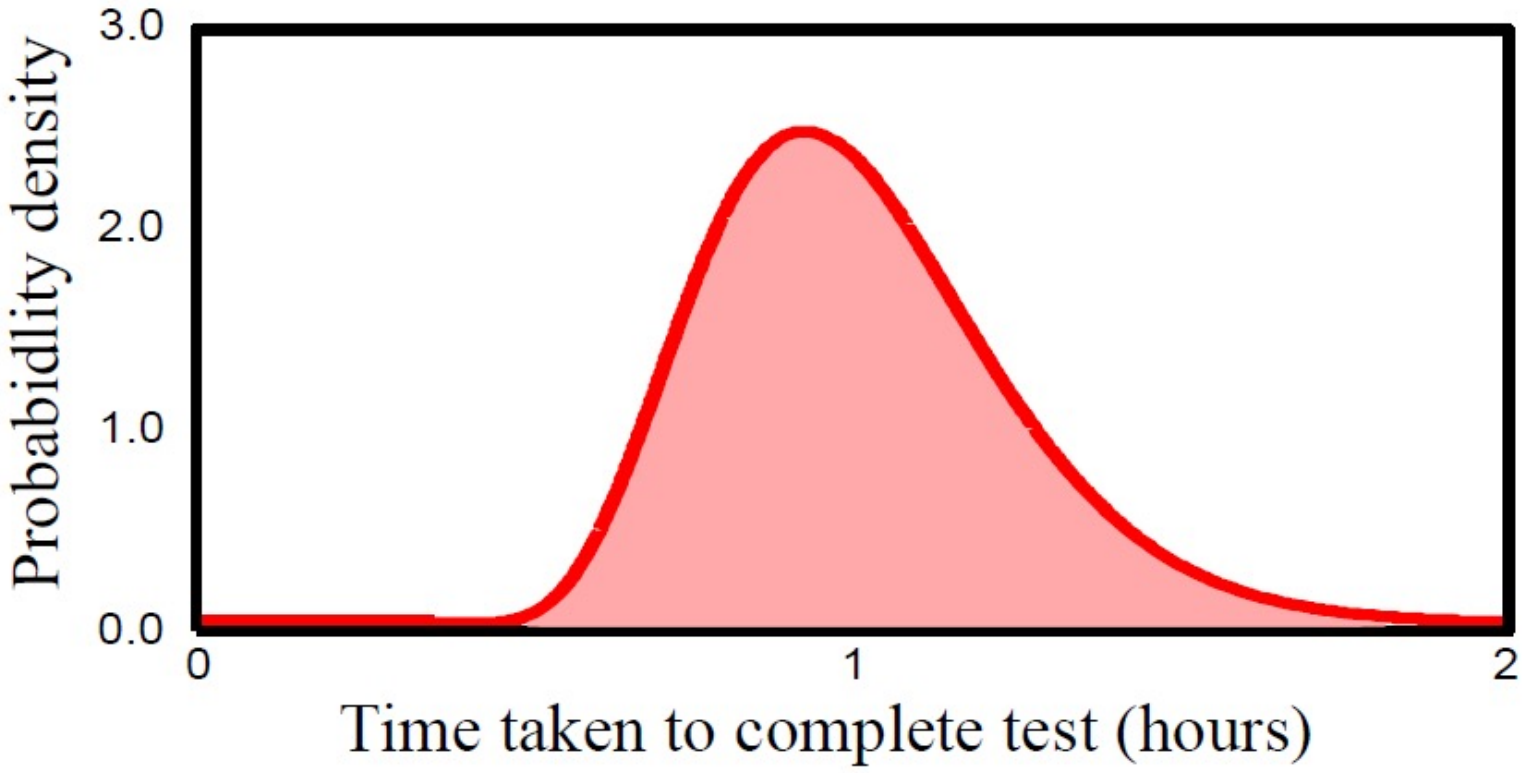
- Take on finite or countably infinite values



$$\sum_x p(x) = 1$$

Adapted from S. Prince

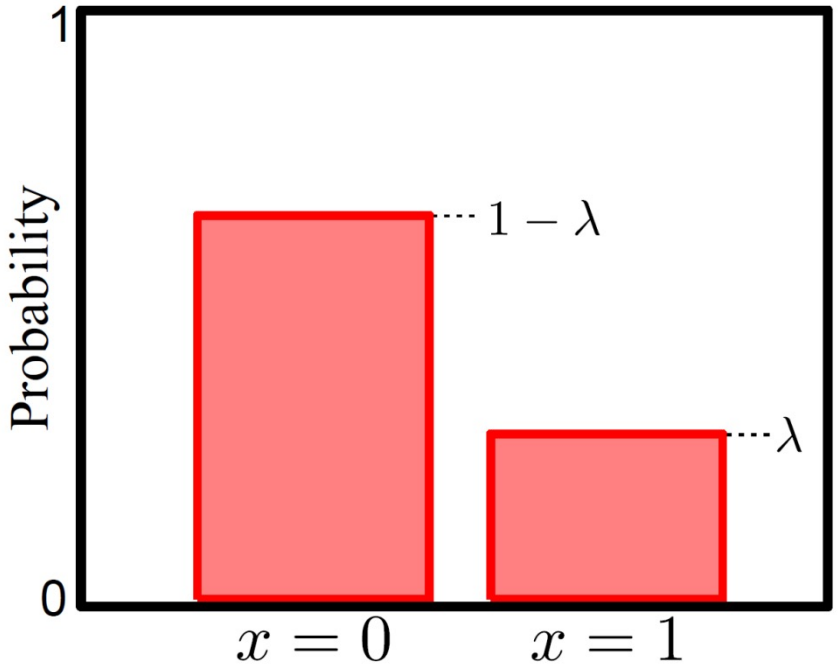
Continuous Random Variable



Famous Discrete Random Variables

- Bernoulli: http://en.wikipedia.org/wiki/Bernoulli_distribution
- Categorical: http://en.wikipedia.org/wiki/Categorical_distribution
- Binomial: http://en.wikipedia.org/wiki/Binomial_distribution
- Geometric: http://en.wikipedia.org/wiki/Geometric_distribution
- Poisson: http://en.wikipedia.org/wiki/Poisson_distribution
- …

Bernoulli Distribution



$$Pr(x=0) = 1-\lambda$$
$$Pr(x=1) = \lambda.$$

or

$$Pr(x) = \lambda^{x}(1-\lambda)^{1-x}$$

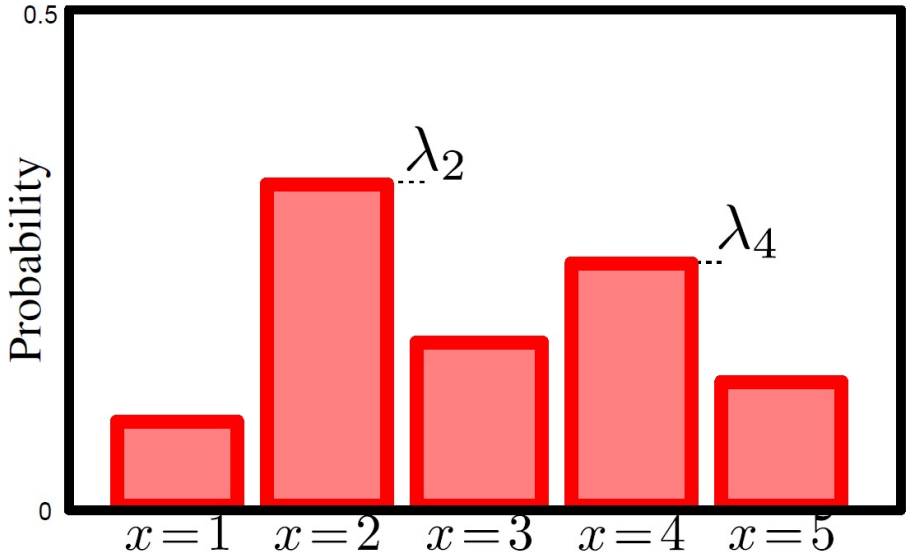
For short we write:

$$p(x) = \text{Ber}(x|\lambda)$$

Bernoulli distribution describes situation where only two possible outcomes x = 0 / x = 1 (e.g. failure/success)

Takes a single parameter $\lambda \in [0,1]$

Categorical Distribution



$$Pr(x = k) = \lambda_k$$

or can think of data as vector with all elements zero except k[th] e.g. $\mathbf{e}_4$ = [0,0,0,1,0]

$$Pr(x = \mathbf{e}_k) = \prod_{j=1}^{K} \lambda_j^{\mathbf{e}_{kj}} = \lambda_k$$

where $\mathbf{e}_{kj}$ is the j-th element of $\mathbf{e}_k$

For short we write: $p(x) = \mathrm{Cat}(x|\lambda)$

Categorical distribution describes situation where K possible outcomes x = 1 , … , x = k , … , x = K.

Takes K parameters $\lambda_k \in [0, 1]$ where $\sum_{k=1}^{K} p(X = k) = \sum_{k=1}^{K} \lambda_k = 1$

$\lambda = \{\lambda_1, \cdots, \lambda_K\}$

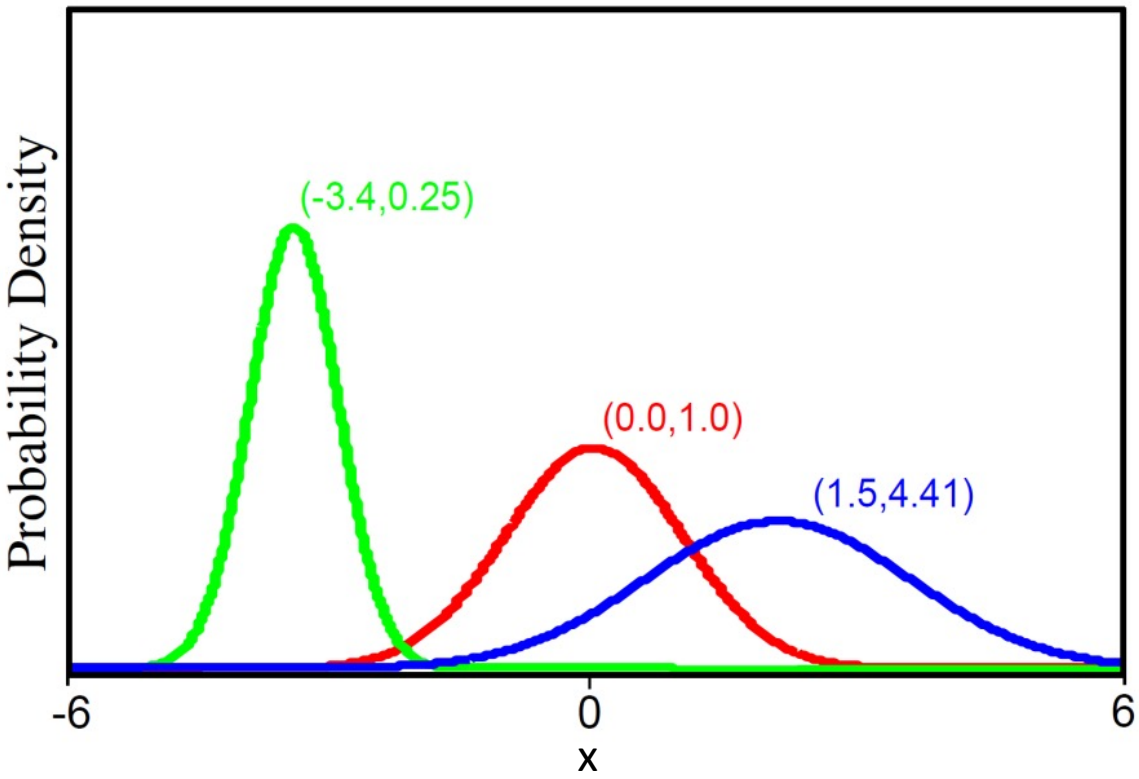
Famous Continuous Random Variables

- Gaussian: http://en.wikipedia.org/wiki/Normal_distribution
- Uniform: http://en.wikipedia.org/wiki/Uniform_distribution_(continuous)
- Exponential: http://en.wikipedia.org/wiki/Exponential_distribution
- Beta: http://en.wikipedia.org/wiki/Beta_distribution
- …

Gaussian/Normal Distribution

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

≤ 0



2 parameters mean μ and variance $\sigma^2>0$

Adapted from S. Prince

Joint Probability

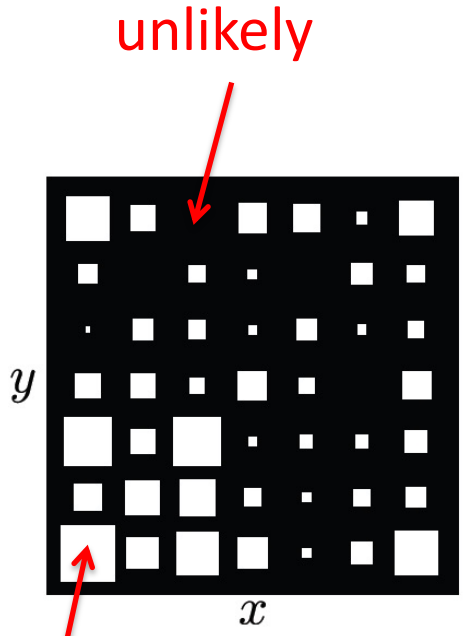
- If we observe two random variables x & y multiple times, then some combinations of outcomes more likely than others

- This information captured by joint probability distribution

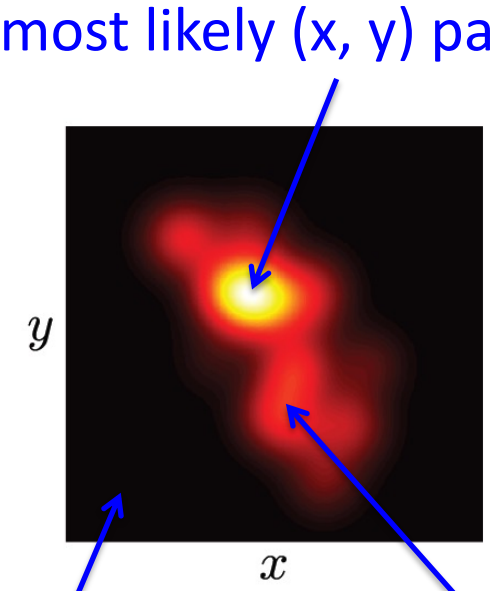
- Written as p(x, y), which is read as "joint probability distribution of x and y"

Adapted from S. Prince

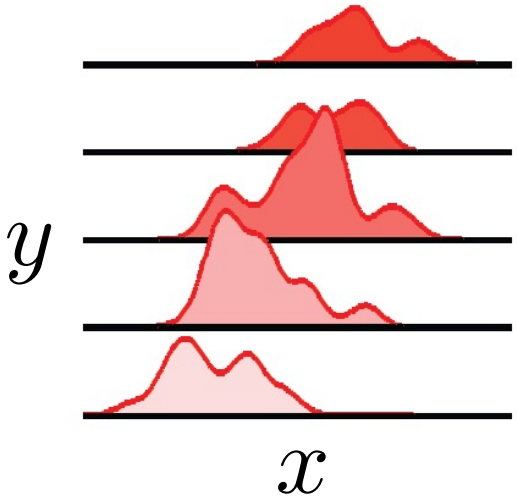
Joint Probability p(x, y)



x & y discrete



x & y continuous



x is continuous,
y is discrete

$$\sum_y \sum_x p(x, y) = 1$$

$$\int_y \int_x p(x, y) dx dy = 1$$

$$\sum_y \int_x p(x, y) dx = 1$$

Adapted from S. Prince

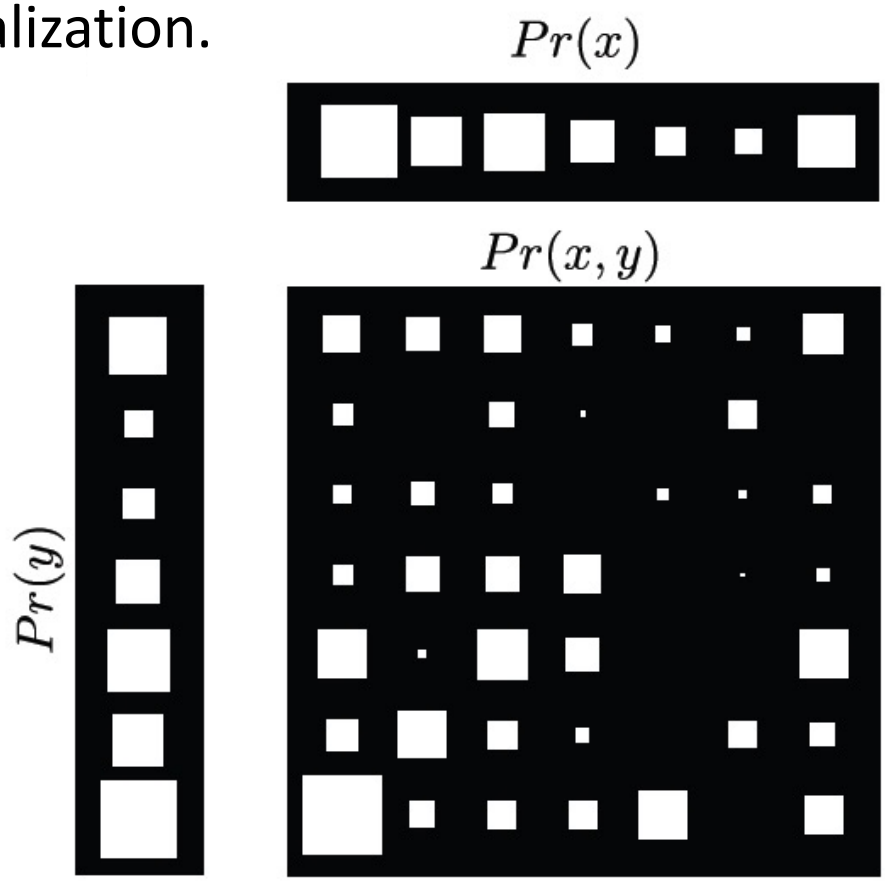
Marginalization / Law of Total Probability

Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variable(s). This is called marginalization.

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

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



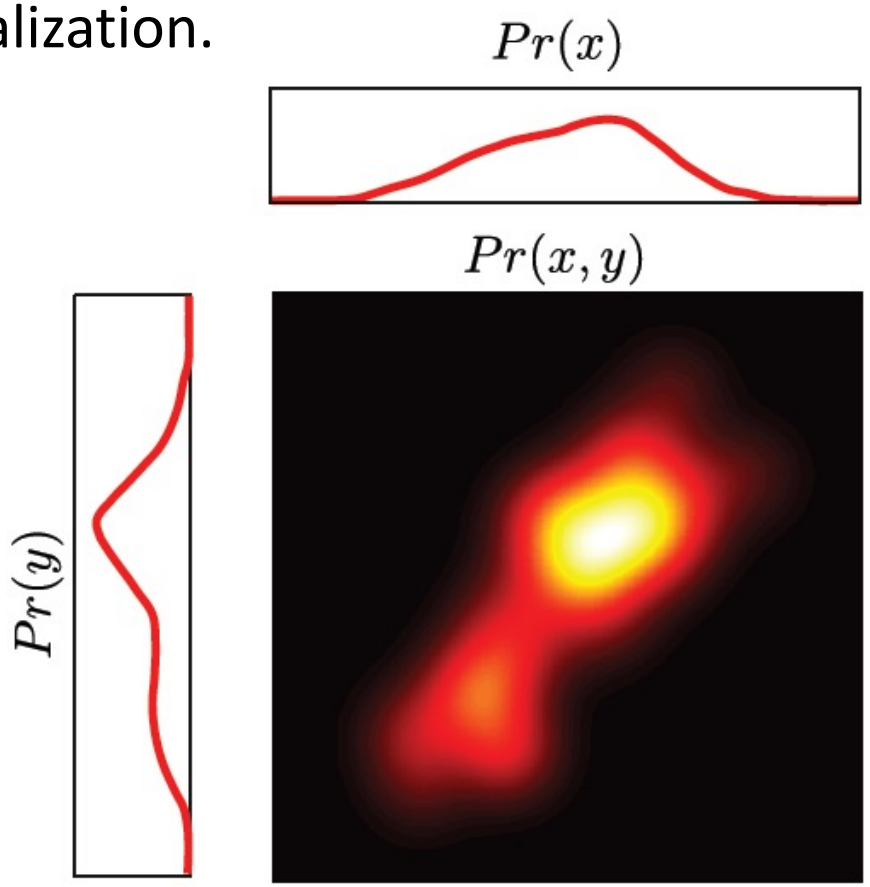
Adapted from S. Prince

Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variable(s). This is called marginalization.

$$p(x) = \int_y p(x, y) dy$$

$$p(y) = \int_x p(x, y) dx$$



Adapted from S. Prince

Marginalization Example

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

p(x, y)

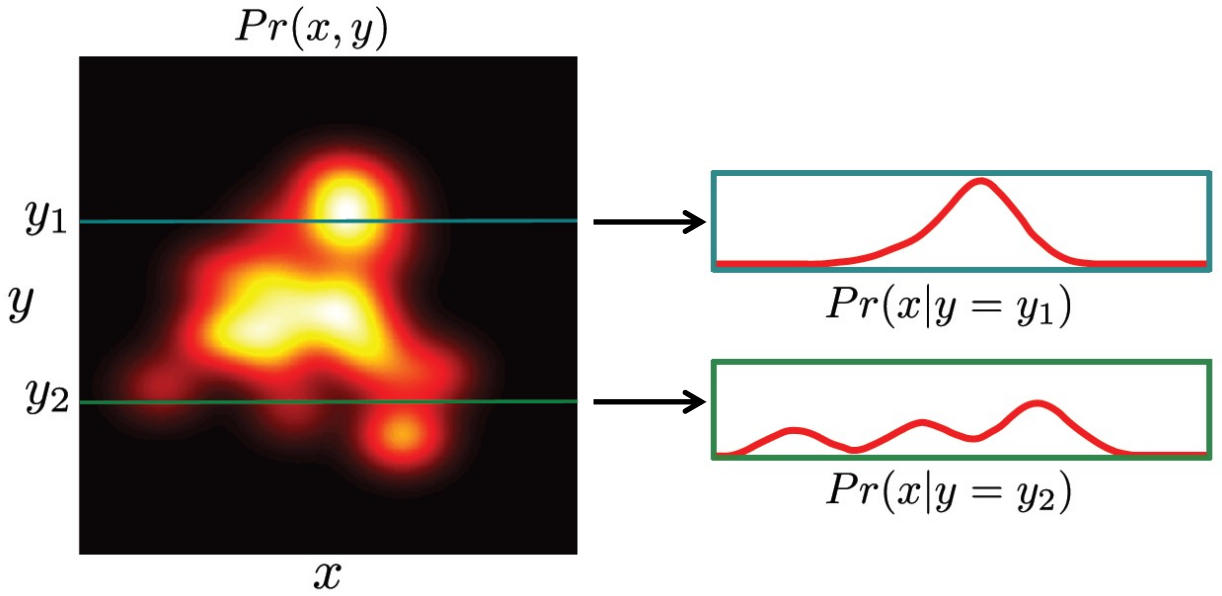
		x		
		0	2.5	p(y)?
	-3	0	1/2	1/2
y	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)?	1/4	3/4	

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

Conditional Probability

Conditional Probability

- Suppose we observe y to be y_1, then p(x | y = y_1) is how likely x will take on various values given this observation
- p(x | y = y_1) read as "conditional probability of X given Y is equal to y_1"



$$\int_x p(x|y = y_1)dx = 1$$

$$\int_x p(x|y = y_2)dx = 1$$

Adapted from S. Prince

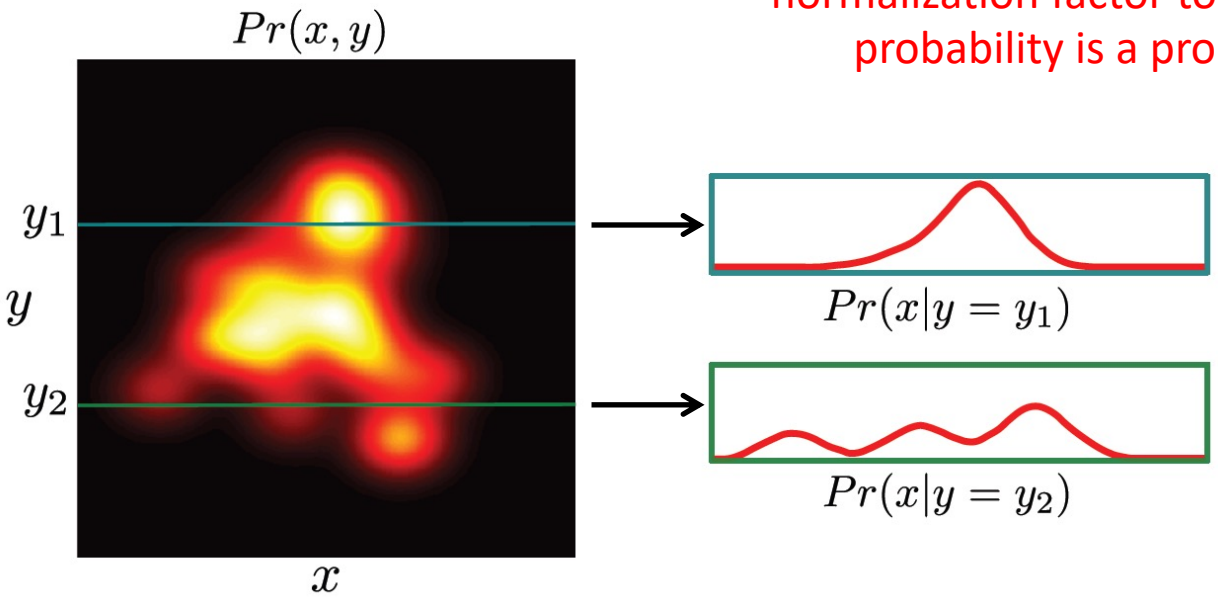
Conditional Probability

- Conditional probability can be computed from joint probability

slice of joint distribution

$$p(x|y=y^*) = \frac{p(x, y=y^*)}{p(y=y^*)} = \frac{p(x, y=y^*)}{\int p(x, y=y^*)dx}$$

normalization factor to ensure conditional probability is a proper distribution



$$\int_x p(x|y=y_1)dx = 1$$

$$\int_x p(x|y=y_2)dx = 1$$

Adapted from S. Prince

Conditional Probability

$$p(x|y=y^*) = \frac{p(x, y=y^*)}{p(y=y^*)} = \frac{p(x, y=y^*)}{\int p(x, y=y^*)dx}$$

- More usually written in compact form

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

- Can be re-arranged to give

$$p(x,y) = p(y)p(x|y)$$
$$p(x,y) = p(x)p(y|x)$$

Adapted from S. Prince

Conditional Probability Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x, y=-1)}{p(y=-1)}$$

$$p(x=0|y=-1) = \frac{p(x=0, y=-1)}{p(y=-1)} = \frac{1/8}{3/8} = \frac{1}{3}$$

$$p(x=2.5|y=-1) =$$

Conditional Probability Example

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x|y=-1) = \frac{p(x, y=-1)}{p(y=-1)}$$

$$p(x=0|y=-1) = \frac{p(x=0, y=-1)}{p(y=-1)} = \frac{1/8}{3/8} = \frac{1}{3}$$

$$p(x=2.5|y=-1) = \frac{p(x=2.5, y=-1)}{p(y=-1)} = \frac{1/4}{3/8} = \frac{2}{3}$$

Bayes' Rule

Deriving Bayes' Rule (y continuous)

From before:

$$p(x,y) = p(y)p(x|y)$$
$$p(x,y) = p(x)p(y|x)$$

Combining:

Equate RHS

$$p(y)p(x|y) = p(x)p(y|x)$$

Re-arranging:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$
$$= \frac{p(y)p(x|y)}{\int p(x,y)dy}$$
$$= \frac{p(y)p(x|y)}{\int p(y)p(x|y)dy}$$

$p(x) = \int p(x,y)dy$

$p(x,y) = p(y)p(x|y)$

Adapted from S. Prince

Deriving Bayes' Rule (y discrete)

From before:

$$p(x,y) = p(y)p(x|y)$$
$$p(x,y) = p(x)p(y|x)$$

Combining:

Equate RHS

$$p(y)p(x|y) = p(x)p(y|x)$$

Re-arranging:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$
$$= \frac{p(y)p(x|y)}{\sum_y p(x,y)}$$
$$= \frac{p(y)p(x|y)}{\sum_y p(y)p(x|y)}$$

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

$$p(x,y) = p(y)p(x|y)$$

Adapted from S. Prince

Bayes' Rule

Prior – what we know about y BEFORE seeing x

Likelihood – propensity for observing a certain value of x given a certain value of y

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} = \frac{p(y)p(x|y)}{\sum_y p(y)p(x|y)}$$

Posterior – what we know about y AFTER seeing x

Evidence – a constant to ensure that the left hand side is a valid distribution

Adapted from S. Prince

Bayes' Rule Example

Example: 40 year old woman doing mammogram

- Let $x = 1$ if mammogram positive, $y = 1$ if breast cancer
- Suppose test is positive, what is cancer probability $p(y = 1|x = 1)$
- Suppose sensitivity $p(x = 1|y = 1) = 0.8$, false positive $p(x = 1|y = 0) = 0.1$, prior $p(y = 1) = 0.004$

$$p(y = 1|x = 1) = \frac{p(y = 1)p(x = 1|y = 1)}{p(y = 1)p(x = 1|y = 1) + p(y = 0)p(x = 1|y = 0)}$$
$$= \frac{0.004 \times 0.8}{0.004 \times 0.8 + 0.996 \times 0.1} = 0.031$$

- US government no longer recommend mammogram for women in 40s

Independence

Independence

- If x & y are independent, then knowing x tells us nothing about y (and vice versa):

$$p(x|y) = p(x)$$
$$p(y|x) = p(y)$$

- If x & y are independent, then joint distribution factorizes into product of marginal distributions:

$$\begin{aligned} p(x,y) &= p(x)p(y|x) \\ &= p(x)p(y) \end{aligned}$$

- Conversely, if joint distribution can be factorized into product of marginal distributions, then x & y are independent

Expectation of One Random Variable

Expectation

Expectation tells us the average (i.e., expected) value of some function f(x) taking into account the distribution of x

Definition:

$$E[f(x)] = \sum_x f(x)p(x)$$

$$E[f(x)] = \int f(x)p(x)dx$$

Adapted from S. Prince

Expectation: Mean and Variance

$$E[f(x)] = \int f(x)p(x)dx$$

- If $f(x) = x$
  - $E[f(x)] = E(x) = \mu_x$, the "mean of x"
  - If we observe x many (infinite) times and average, we get μ_x
- If $f(x) = (x - \mu_x)^2$
  - $E[f(x)] = E[(x - \mu_x)^2] = \sigma_x^2$
  - $\sigma_x^2 = Var(x)$ called "variance"; σ_x called "standard deviation"
  - If we observe x many (infinite) times and average square of difference between each observation and μ_x, we get σ_x^2
  - Measure how likely x is going to be far away from mean

Expectation Example: Mean

		x = 0	x = 2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E[f(x)] = \sum_x f(x)p(x)$$

$$E(x) = \sum_x xp(x)$$

$$= 0 \times 1/4 + 2.5 \times 3/4$$

$$= 0 + 1.875$$

$$= 1.875$$

Expectation Example: Mean

		x: 0	x: 2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E[f(x)] = \sum_x f(x)p(x)$$

$$E(x) = \sum_x xp(x)$$

$$= 0 \times 1/4 + 2.5 \times 3/4$$

$$= 0 + 1.875$$

$$= 1.875$$

$$E(y) = \sum_y yp(y)$$

$$= -3 \times 1/2 + (-1) \times 3/8 + 2 \times 1/8$$

$$= -3/2 - 3/8 + 1/4$$

$$= -1.625$$

Expectation Example: Variance

		x = 0	x = 2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E[f(x)] = \sum_x f(x)p(x)$$

$$E(x) = \sum_x xp(x)$$

$$= 0 \times 1/4 + 2.5 \times 3/4$$

$$= 0 + 1.875$$

$$= 1.875$$

$$E[f(x)] = \sum_x f(x)p(x)$$

$$E[(x - \mu_x)^2] = \sum_x (x - \mu_x)^2 p(x)$$

$$= (0 - 1.875)^2 \times 1/4 + (2.5 - 1.875)^2 \times 3/4$$

$$= 1.171875$$

Expectation Example: Variance

		x = 0	x = 2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E[f(x)] = \sum_x f(x)p(x)$$

$$E(x) = \sum_x xp(x)$$

$$= 0 \times 1/4 + 2.5 \times 3/4$$

$$= 0 + 1.875$$

$$= 1.875$$

$$E[f(x)] = \sum_x f(x)p(x)$$

$$E[(x - \mu_x)^2] = \sum_x (x - \mu_x)^2 p(x)$$

$$= (0 - 1.875)^2 \times 1/4 + (2.5 - 1.875)^2 \times 3/4$$

$$= 1.171875$$

Expectation of Two Random Variables

Expectation for X, Y

- Expectation tells us the expected or average value of some function $f(x, y)$ taking into account $p(x, y)$

$$E[f(x,y)] = \int \int f(x,y)p(x,y)dxdy$$

- Special case: $f(x, y) = (x–\mu_x)(y–\mu_y)$
  - $E[f(x, y)] = E[(x–\mu_x)(y–\mu_y)] = Cov(x, y)$, the covariance of x and y
  - Measure how much two variables change together
  - $Cov(x, y)$ positive whenever $x > \mu_x$, then $y > \mu_y$ on average (& vice versa)
  - $Cov(x, y)$ negative whenever $x > \mu_x$, then $y < \mu_y$ on average (& vice versa)

Expectation Example: Covariance

		x: 0	x: 2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$
$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$E[(x-\mu_x)(y-\mu_y)]$$
$$= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y)$$
$$= (0-1.875)(-3+1.625) \times 0 + (2.5-1.875)(-3+1.625) \times 1/2 +$$
$$(0-1.875)(-1+1.625) \times 1/8 + (2.5-1.875)(-1+1.625) \times 1/4 +$$
$$(0-1.875)(2+1.625) \times 1/8 + (2.5-1.875)(2+1.625) \times 0$$
$$= -1.328125$$

Expectation Example: Covariance

		x = 0	x = 2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$
$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$\begin{aligned}
&E[(x-\mu_x)(y-\mu_y)] \\
&= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y) \\
&= (0-1.875)(-3+1.625)\times 0 + (2.5-1.875)(-3+1.625)\times 1/2 + \\
&\quad (0-1.875)(-1+1.625)\times 1/8 + (2.5-1.875)(-1+1.625)\times 1/4 + \\
&\quad (0-1.875)(2+1.625)\times 1/8 + (2.5-1.875)(2+1.625)\times 0 \\
&= -1.328125
\end{aligned}$$

Expectation Example: Covariance

		x = 0	x = 2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$
$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$E[(x-\mu_x)(y-\mu_y)]$$
$$= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y)$$
$$= (0-1.875)(-3+1.625) \times 0 + (2.5-1.875)(-3+1.625) \times 1/2 +$$
$$(0-1.875)(-1+1.625) \times 1/8 + (2.5-1.875)(-1+1.625) \times 1/4 +$$
$$(0-1.875)(2+1.625) \times 1/8 + (2.5-1.875)(2+1.625) \times 0$$
$$= -1.328125$$

Expectation Example: Covariance

		x		
		0	2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$
$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$\begin{aligned}
&E[(x-\mu_x)(y-\mu_y)] \\
&= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y) \\
&= (0-1.875)(-3+1.625)\times 0 + (2.5-1.875)(-3+1.625)\times 1/2 + \\
&\quad (0-1.875)(-1+1.625)\times 1/8 + (2.5-1.875)(-1+1.625)\times 1/4 + \\
&\quad (0-1.875)(2+1.625)\times 1/8 + (2.5-1.875)(2+1.625)\times 0 \\
&= -1.328125
\end{aligned}$$

Expectation Example: Covariance

		x = 0	x = 2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$
$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$E[(x-\mu_x)(y-\mu_y)]$$
$$= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y)$$
$$= (0-1.875)(-3+1.625) \times 0 + (2.5-1.875)(-3+1.625) \times 1/2 +$$
$$(0-1.875)(-1+1.625) \times 1/8 + (2.5-1.875)(-1+1.625) \times 1/4 +$$
$$(0-1.875)(2+1.625) \times 1/8 + (2.5-1.875)(2+1.625) \times 0$$
$$= -1.328125$$

Expectation Example: Covariance

		x = 0	x = 2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$E(x) = \mu_x = 1.875$$
$$E(y) = \mu_y = -1.625$$

$$E[f(x,y)] = \sum_x \sum_y f(x,y)p(x,y)$$

$$E[(x-\mu_x)(y-\mu_y)]$$
$$= \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y)$$
$$= (0-1.875)(-3+1.625) \times 0 + (2.5-1.875)(-3+1.625) \times 1/2 +$$
$$(0-1.875)(-1+1.625) \times 1/8 + (2.5-1.875)(-1+1.625) \times 1/4 +$$
$$(0-1.875)(2+1.625) \times 1/8 + (2.5-1.875)(2+1.625) \times 0$$
$$= -1.328125$$

Rules of Expectation

1. Expectation of a non-random constant: $E(c) = c$
2. Assuming a and b are non-random constants, then
   - $E[af(x) + bg(x)] = aE(f(x)) + bE(g(x))$
   - This is called "linearity of expectation"
3. $Cov(x, y) = E(xy) - E(x)E(y)$
   - Therefore $Var(x) \equiv Cov(x, x) = E(x^2) - (E(x))^2$
4. If x and y are independent, then $E[f(x)g(y)] = E[f(x)]E[g(y)]$
   - If x and y are independent, $Cov(x, y) = 0$
   - However $Cov(x, y) = 0 \nRightarrow$ independence

Conditional Expectation

Conditional Expectation

- Remember $p(x|y)$ is conditional probability of x given y?
- Conditional expectation:

$$\begin{aligned} E[f(x,y)|y] &\triangleq E_{p(x|y)}[f(x,y)] \\ &\triangleq \sum_x f(x,y)p(x|y) \end{aligned}$$

- Read as "expected value of $f(x,y)$ given y"
- Conditional expectation tells us average value of $f(x,y)$ taking into account $p(x|y)$

Conditional Expectation

- Remember $p(x|y)$ is conditional probability of x given y?
- Conditional expectation:

$$E[f(x,y)|y] \triangleq E_{p(x|y)}[f(x,y)]$$
$$\triangleq \int_x f(x,y)p(x|y)dx$$

- Read as "expected value of $f(x,y)$ given y"
- Conditional expectation tells us average value of $f(x,y)$ taking into account $p(x|y)$

Conditional Expectation Example

		x: 0	x: 2.5	p(y)
y	-3	0	1/2	1/2
	-1	1/8	1/4	3/8
	2	1/8	0	1/8
	p(x)	1/4	3/4	

$$p(x = 0|y = -1) = \frac{1}{3}$$

$$p(x = 2.5|y = -1) = \frac{2}{3}$$

$$E(x) = 1.875$$

$$E_{p(x|y=-1)}[x] = \sum_x x p(x|y = -1) \quad \leftarrow \quad \sum_x f(x, y) p(x|y)$$

$$= 0 \times p(x = 0|y = -1) + 2.5 \times p(x = 2.5|y = -1)$$

$$= 0 \times 1/3 + 2.5 \times 2/3$$

$$= 1.667$$

N Random Variables

N random variables (aka random vector)

- Have focused on 2 random variables x and y
- In real applications, usually more than 2 variables (e.g., photo has > 1M pixels)
- If we observe $x_1, x_2, \ldots, x_N$ multiple times, some combinations of outcomes more likely than others
- This information captured by joint probability distribution function
- Written as $p(x_1, x_2, \ldots, x_N)$, read as probability distribution of x_1 to x_N
- If $x_1, x_2, \ldots, x_N$ continuous, then p refers to joint probability distribution function (pdf). If discrete, then refers to joint probability mass function (pmf)
- Many properties for two random variables generalize naturally to more variables

Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variables

$$Pr(x) = \int Pr(x,y)\, dy$$

$$Pr(y) = \int Pr(x,y)\, dx$$

Works in higher dimensions as well – leaves joint distribution between whatever variables are left

$$Pr(x,y) = \sum_{w} \int Pr(w,x,y,z)\, dz$$

Conditional Probability

- Two variables

$$p(x, y) = p(x)p(y|x)$$

- Three variables

$$p(a, b, c) = p(a)p(b, c|a) = p(a)p(b|a)p(c|a, b)$$

- N variables

$$\begin{aligned} p(x_1, \cdots, x_N) &= p(x_1)p(x_2, \cdots, x_N|x_1) \\ &= p(x_1)p(x_2|x_1)p(x_3, \cdots, x_N|x_1, x_2) \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_N|x_1, \cdots, x_{N-1}) \end{aligned}$$

Independence

- If $x_1, \cdots, x_N$ are independent, then knowing any subset of x's tells us nothing about the remaining x's
- If $x_1, \cdots, x_N$ are independent if and only if the joint distribution factorizes into product of marginal distributions:

$$p(x_1, \cdots, x_N) = p(x_1)p(x_2)\cdots p(x_N) \stackrel{\triangle}{=} \prod_{n=1}^{N} p(x_n)$$

- $x_1, \cdots, x_N$ are independently and identically distributed (i.i.d.) if they are independent and $p(x_1) = p(x_2) = \cdots = p(x_N)$

Conditional Independence

- x_1 and x_2 are conditionally independent given x_3 if and only if

$$p(x_1, x_2|x_3) = p(x_1|x_3)p(x_2|x_3)$$

Knowing x_2 tells us nothing about x_1 (and vice versa) if we already know x_3

Special Case I

- If x_1 and x_2 are independent, it does NOT imply x_1 and x_2 are conditionally independent given x_3
- Example: Let x = 1 if Thomas comes to class wet (0 otherwise). Let y = 1 if raining (0 otherwise). Let z = 1 if Thomas involved in water fight (0 otherwise)
  - Then y and z are independent (presumably)
  - But y and z are not conditionally independent given x
  - Because suppose I come to class wet. Then knowing it's not raining would suggest I was in a water fight

Special Case II

- If x_1 and x_2 are conditionally independent given x_3, it does NOT imply x_1 and x_2 are independent
- Example: Toss coin 99 times. Let q = probability of head. Let x_n = 1 if n-th coin toss = head and x_n= 0 if n-th coin toss = tail
  - x_1, x_2, … x_{99} are independent conditioned on knowing q (e.g., q = 0.7)
  - If q is NOT known, then x_1, x_2, … x_{99} are not independent
  - Imagine if all 99 coin tosses are equal to head, what would you guess about the 100th coin toss?
- Probably confusing because in your previous statistics class, you might have seen equation like $p(HHT) = q^2(1-q)$
  - This cannot be right since q does not appear on left hand side? How can it appear on right hand side?
  - Instead it should actually be $p(HHT \mid q) = q^2(1-q)$. The term "| q" is implicit and is often dropped to reduce clutter

Summary

- Discrete & continuous random variables
- Probability distribution function / probability mass function
- Joint distributions of N random variables
- Marginalization / Law of Total Probability
- Conditional probability
- Bayes' Rule
- Independence, Conditional Independence
- Expectation, conditional expectation

Optional Reading II

- These notes are from Chapters 2 and 3 of SP (free download: www.computervisionmodels.com)
- Review of probability & linear algebra by Dr. Tam Nguyen (download from LuminNUS)
- Textbooks
  - Appendix A of D&H
  - Chapter 2 of KM (beware of typos)

Additional Material

Rules of Expectations

- E[af(x) + bg(x)] = aE(f(x)) + bE(g(x))

$$
\begin{aligned}
E(af(x) + bg(x)) &= \sum_x (af(x) + bg(x))p(x) \\
&= a\sum_x f(x)p(x) + b\sum_x g(x)p(x) \\
&= aE(f(x)) + bE(g(x))
\end{aligned}
$$

Rules of Expectations

- Cov(x,y) = E(xy)-E(x)E(y)

$$
\begin{aligned}
&Cov(x,y) \\
&= E[(x - E(x))(y - E(y))] \\
&= E[(xy - xE(y) - yE(x) + E(x)E(y))] \\
&= E(xy) - E(xE(y)) - E(E(x)y) + E(E(x)E(y)) \\
&= E(xy) - E(x)E(y) - E(x)E(y) + E(x)E(y) \\
&= E(xy) - E(x)E(y)
\end{aligned}
$$

Expand the terms

Linearity of Expectation

E(x) & E(y) are constants

Rules of Expectations

- If x & y are independent, then E[f(x)g(y)] = E[f(x)]E[g(y)]

$$E(f(x)g(y)) = \sum_x \sum_y f(x)g(y)p(x,y)$$

$$= \sum_x \sum_y f(x)g(y)p(x)p(y)$$

x & y are indepdent

$$= \sum_x f(x)p(x) \sum_y g(y)p(y)$$

$$= E(f(x))E(g(y))$$