Discussion Week 4

Overview

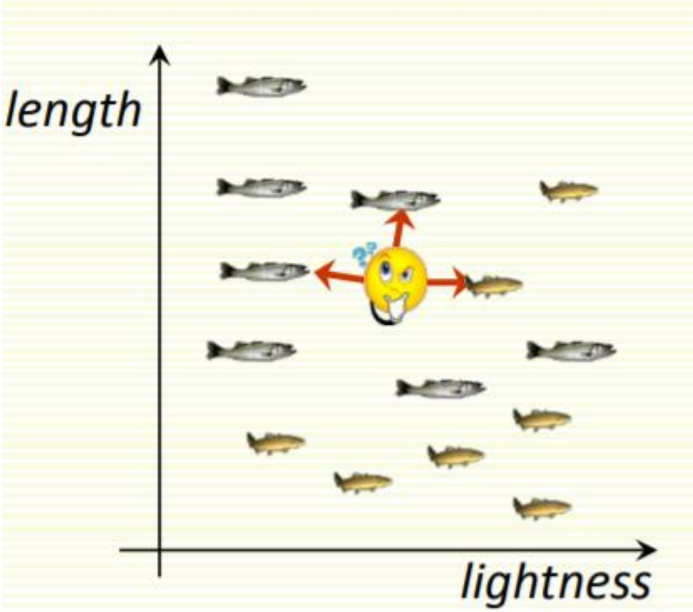
- KNN
- Similarity Metrics
- Classification Evaluation

KNN

KNN Algorithm

- classify an unknown example with the most common class among k closest examples
  - “tell me who your neighbors are, and I’ll tell you who you are”

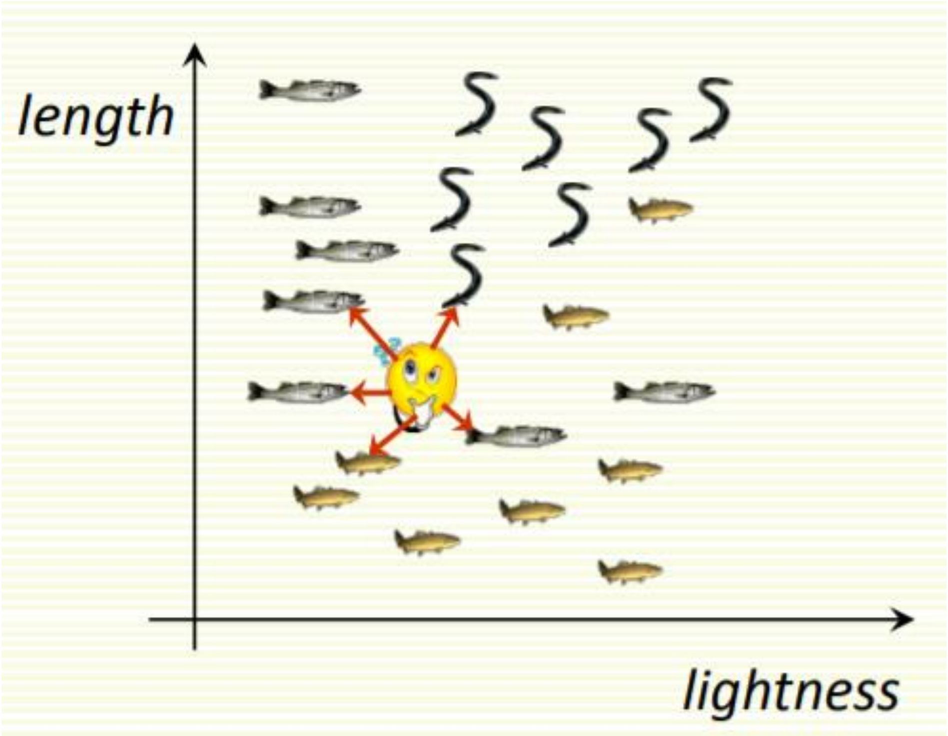
- Example
  - k = 3
  - 2 sea bass, 1 salmon
  - Classify as sea bass



KNN: Multiple Classes

- Easy to implement for multiple classes

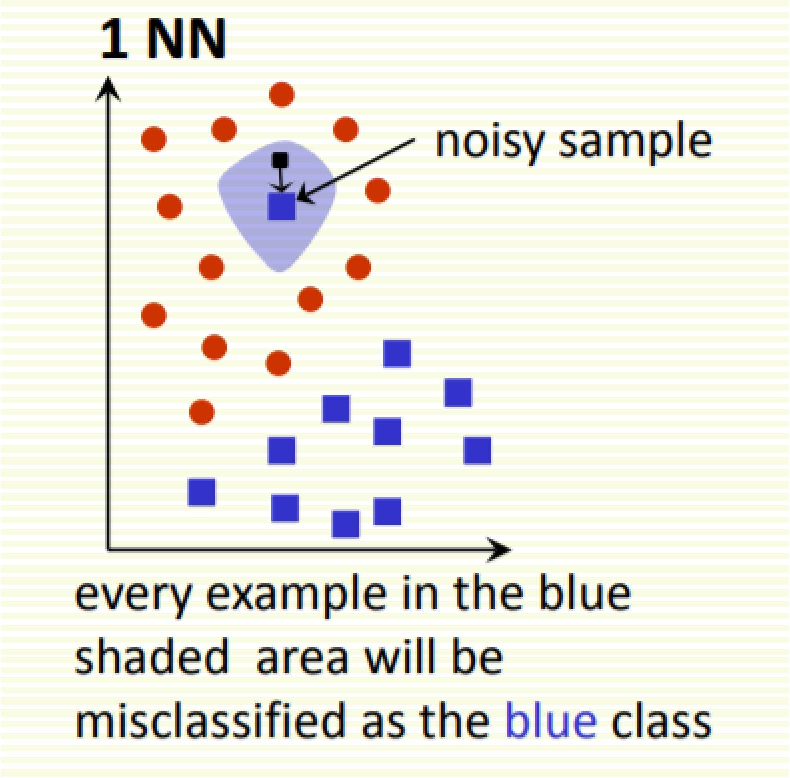
- Example for k = 5
  - 3 fish species: salmon, sea bass, eel
  - 3 sea bass, 1 eel, 1 salmon ⇒ classify as sea bass

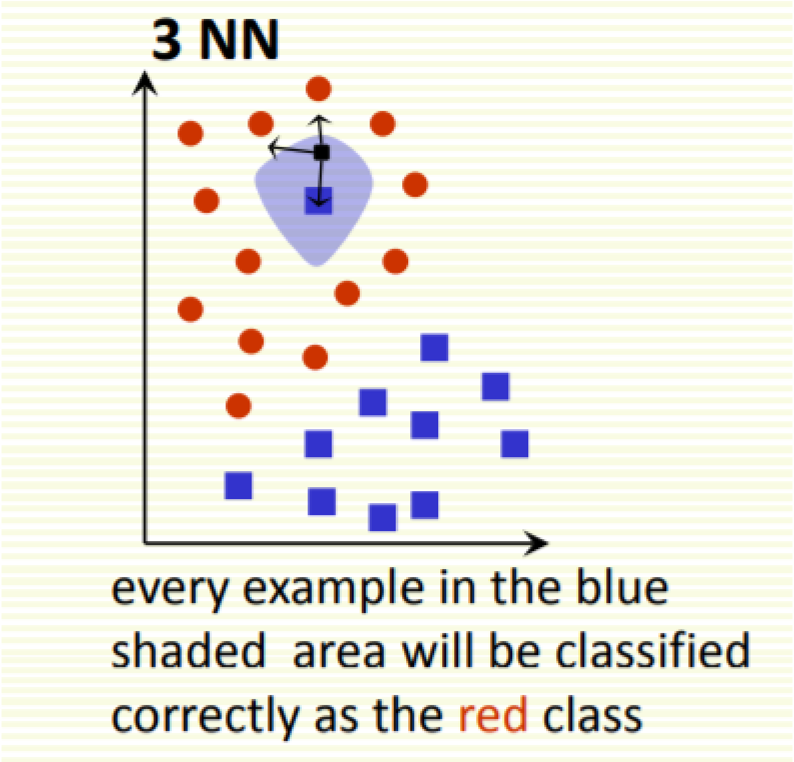


KNN: How to Choose k?

- In theory, if infinite number of samples available, the larger the k, the better is classification
- Caveat: all k neighbors have to be close
  - Possible when infinite # samples available
  - Impossible in practice since # samples is finite
- Should we "tune" k on training data?
  - Issue: overfitting
- k = 1 is efficient, but sensitive to "noise"

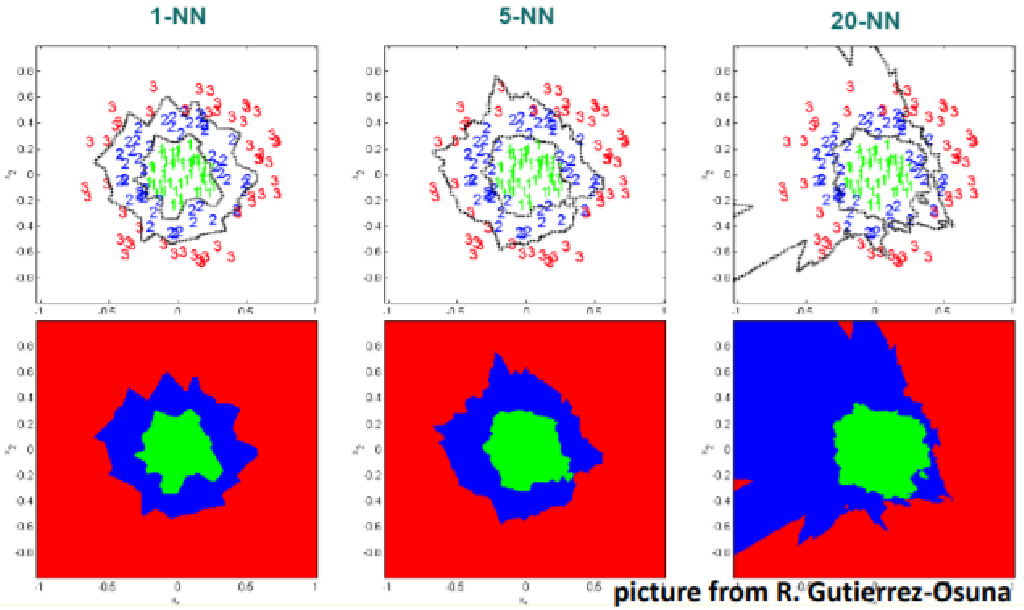
KNN: How to Choose k?





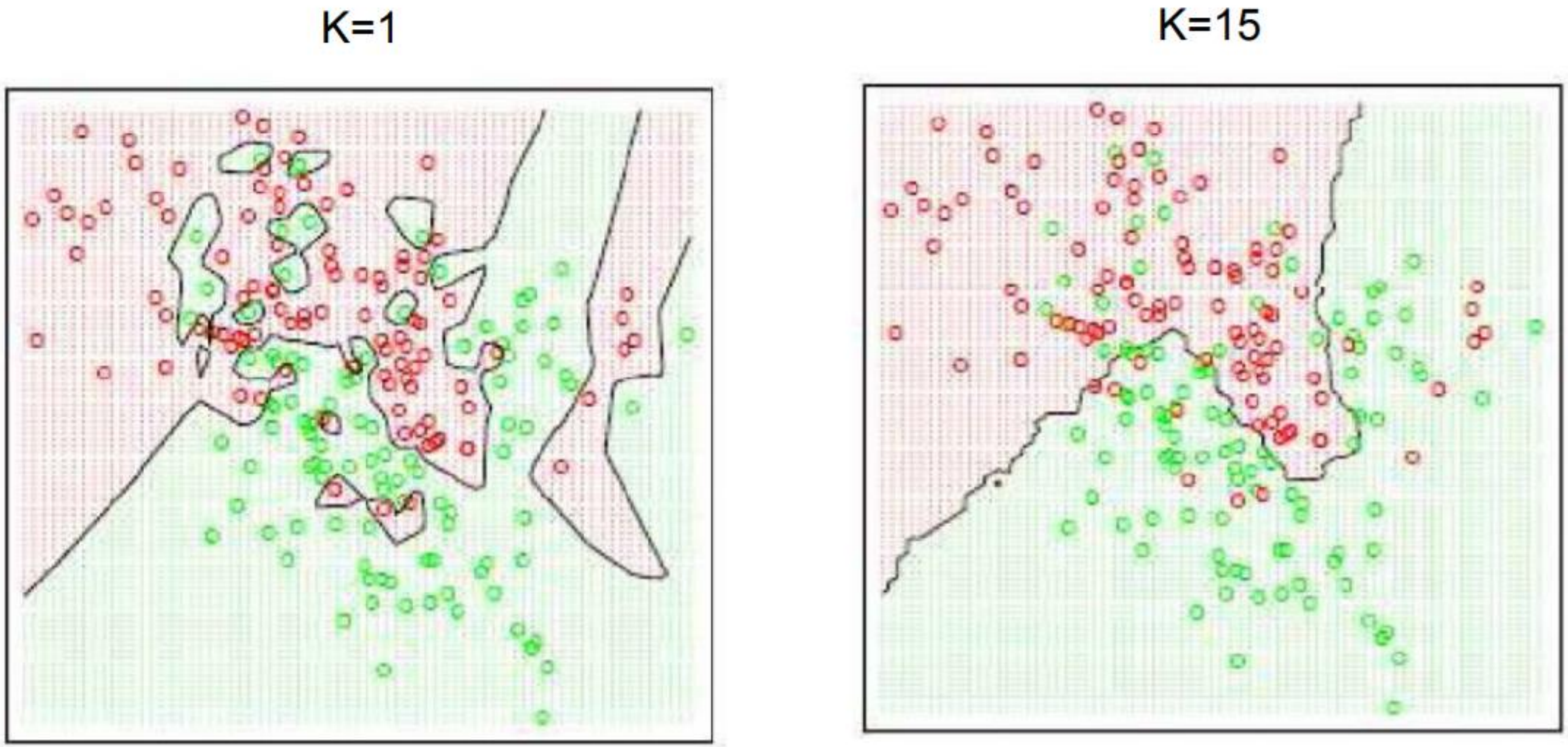
KNN: How to Choose k?

- Larger k gives smoother boundaries, better for generalization
  - Only if locality is preserved
  - k too large => end up looking at samples too far away not from the same class
- Can choose k through cross-validation



picture from R. Gutierrez-Osuna

KNN: How to Choose k?

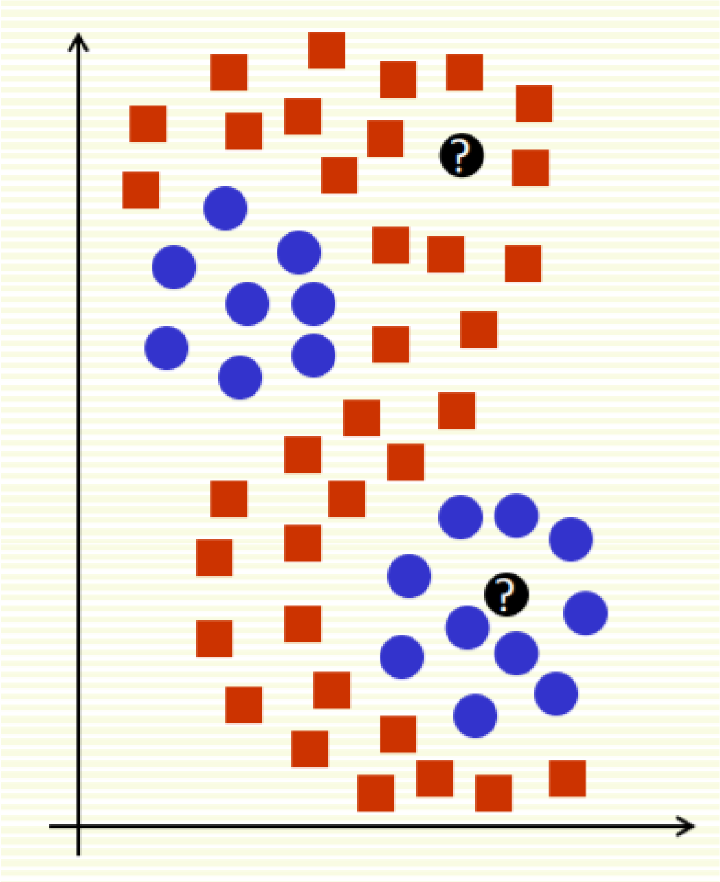


Figures from Hastie, Tibshirani and Friedman (Elements of Statistical Learning)

KNN: Multi-Modal Distributions

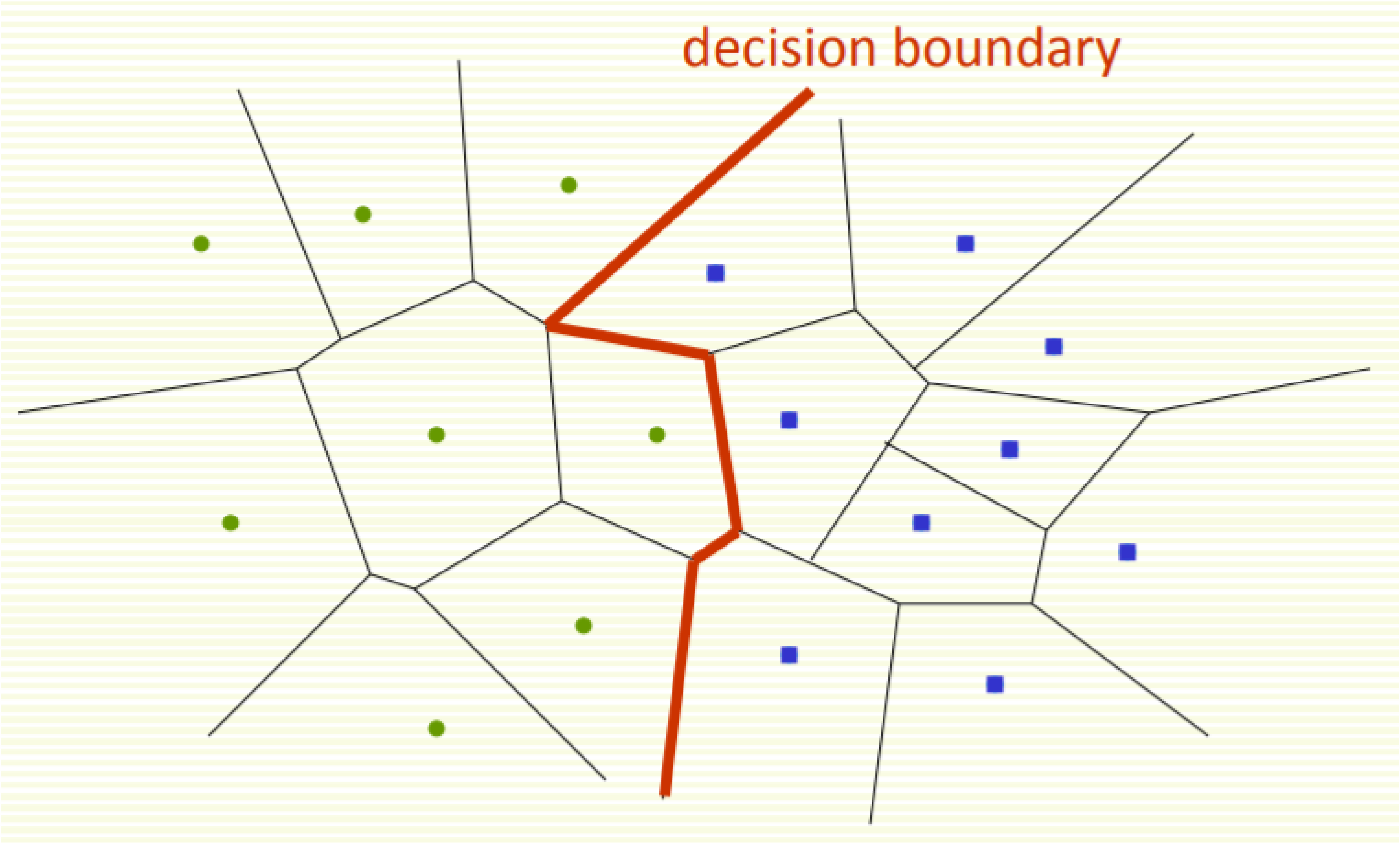
- Many classification models would not work for this 2 class classification problem

- Nearest neighbors will do reasonably well, provided we have a lot of samples



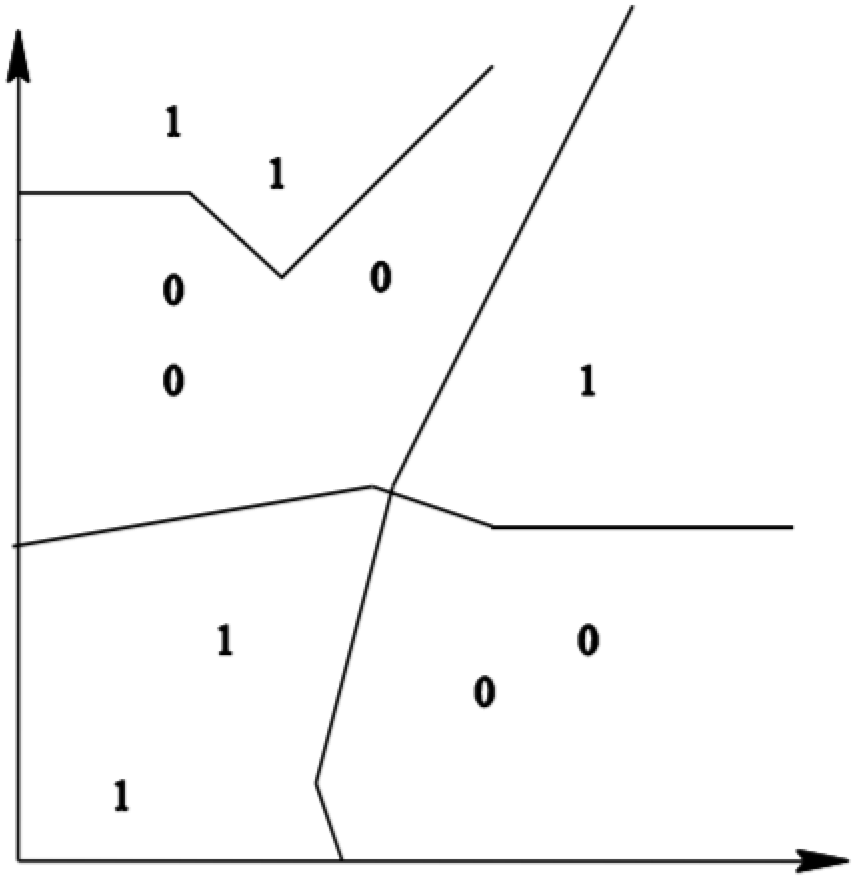
Decision Boundaries

- Voronoi diagram is useful for visualization



Decision Boundaries

- Decision boundaries are formed by a subset of the Voronoi diagram of the training data
- Each line segment is equidistant between two points of opposite class
- The more examples that are stored, the more fragmented and complex the decision boundaries can become.



KNN Selection of Distance

- So far we assumed we use Euclidian Distance to find the nearest neighbor:

$$D(a, b) = \sqrt{\sum_k (a_k - b_k)^2}$$

- Euclidean distance treats each feature as equally important
- However some features (dimensions) may be much more discriminative than other features

KNN Distance Selection: Extreme Example

- feature 1 gives the correct class: 1 or 2
- feature 2 gives irrelevant number from 100 to 200
- dataset: [1   150], [2   110]
- classify [1   100]

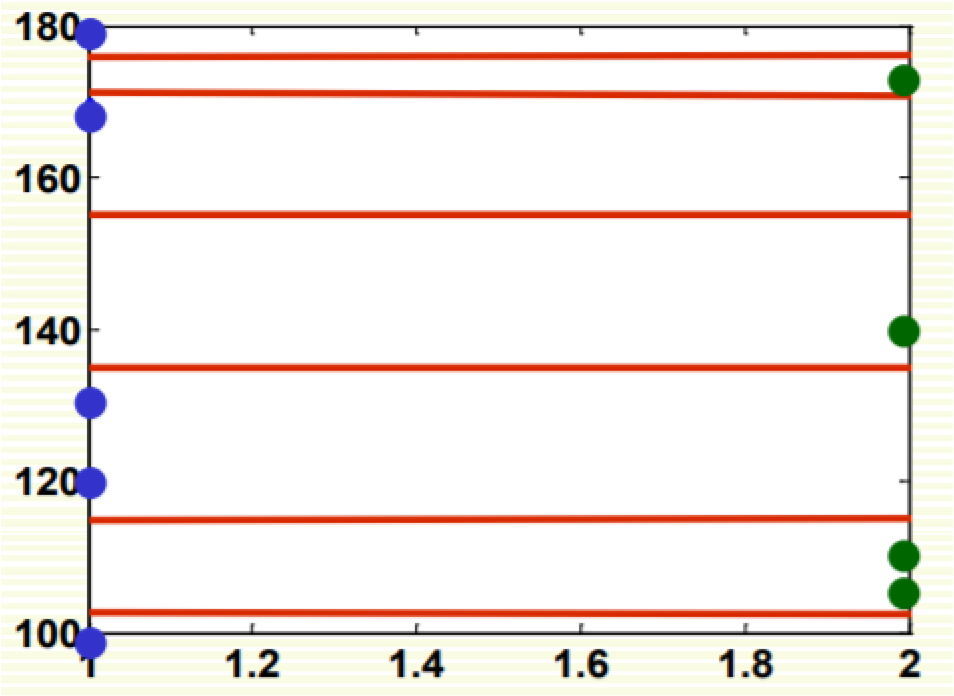
$$D\left(\begin{bmatrix}1\\100\end{bmatrix},\begin{bmatrix}1\\150\end{bmatrix}\right) = \sqrt{(1-1)^2 + (100-150)^2} = 50$$

$$D\left(\begin{bmatrix}1\\100\end{bmatrix},\begin{bmatrix}2\\110\end{bmatrix}\right) = \sqrt{(1-2)^2 + (100-110)^2} = 10.5$$

- [1   100] is misclassified!
- The denser the samples, the less of this problem
- But we rarely have samples dense enough

KNN Distance Selection: Extreme Example

- Decision boundary is in red, and is really wrong because
  - feature 1 is discriminative, but it’s scale is small
  - feature 2 gives no class information but its scale is large, it dominates distance calculation



KNN: Feature Normalization

- Notice that 2 features are on different scales
- First feature takes values between 1 or 2
- Second feature takes values between 100 to 200
- Idea: normalize features to be on the same scale
- Different normalization approaches
- Linearly scale the range of each feature to be, say, in range [0,1]

$$f_{new} = \frac{f_{\text{old}} - f_{\text{old}}^{\text{min}}}{f_{\text{old}}^{max} - f_{\text{old}}^{\text{min}}}$$

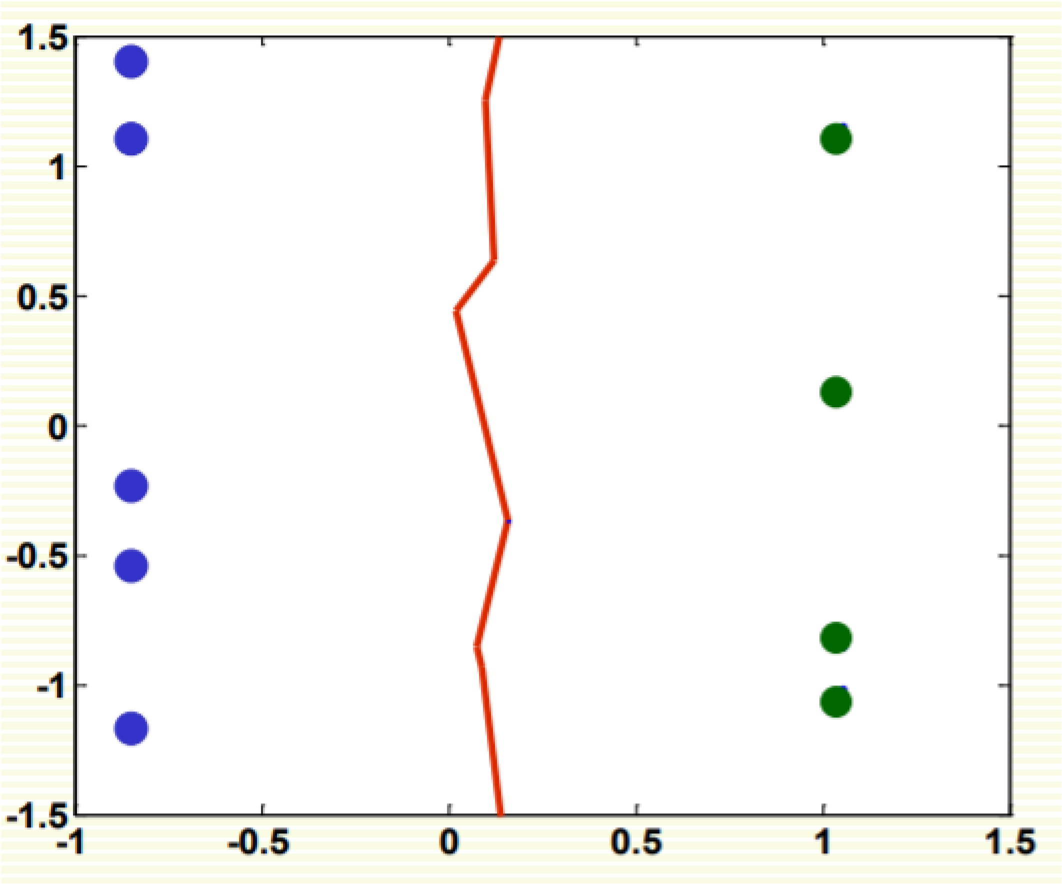
KNN: Feature Normalization

- Linearly scale to 0 mean variance 1
- If ***z*** is a random variable of mean μ and variance σ^2, then (***z*** - μ)/σ has mean 0 and variance 1
- For each feature f let the new rescaled feature be

$$f_{new} = \frac{f_{\mathrm{old}} - \mu}{\sigma}$$

- Let us apply this normalization to previous example

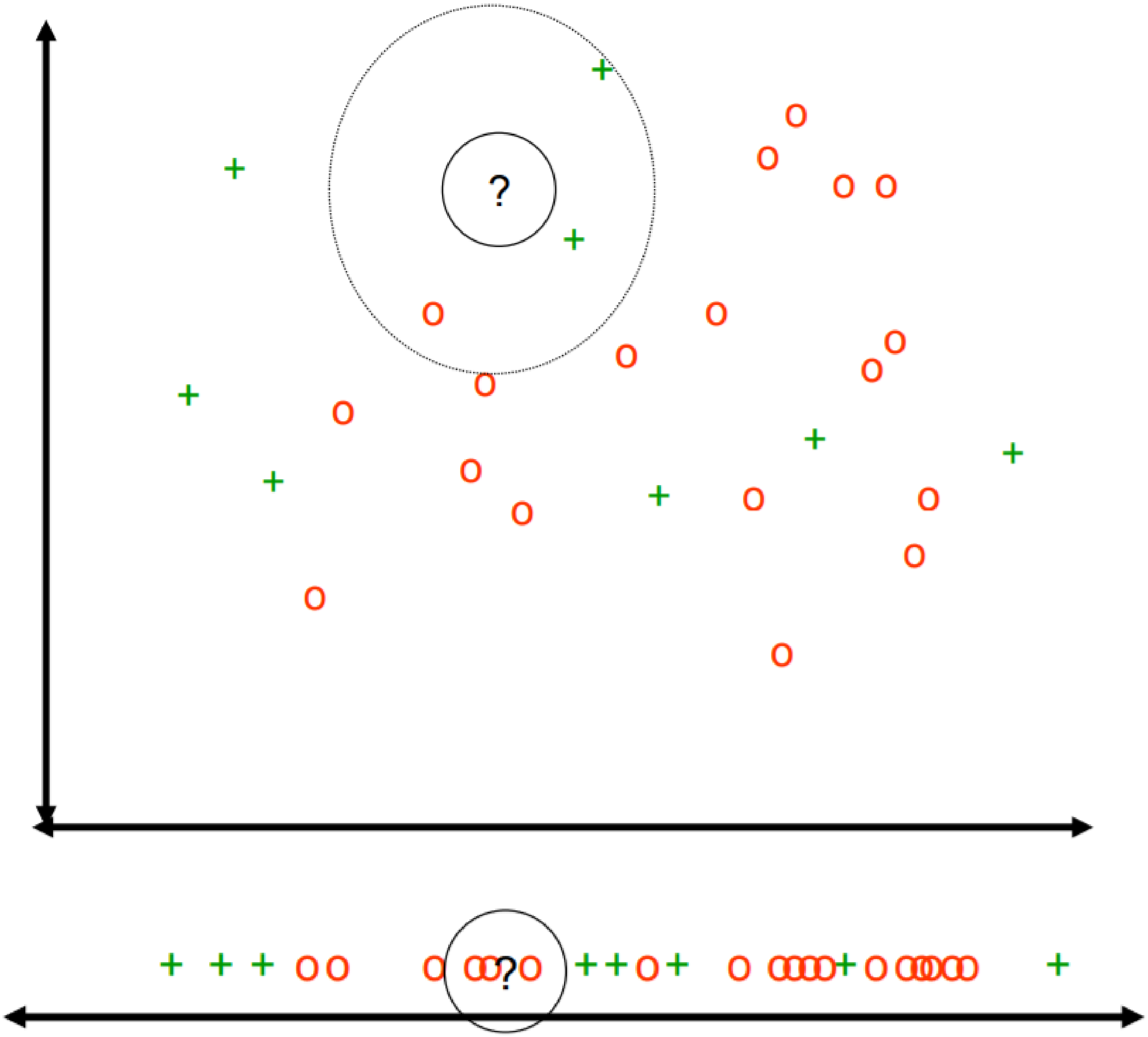
KNN: Feature Normalization



KNN: Selection of Distance

- Feature normalization does not help in high dimensional spaces if most features are irrelevant
- $D(a,b) = \sqrt{\sum_k (a_k - b_k)^2} = \sqrt{\underbrace{\sum_i (a_i - b_i)^2}_{\text{Discriminative features}} + \underbrace{\sum_j \left(a_j - b_j\right)^2}_{\text{Noisy features}}}$

- If the number of useful features is smaller than the number of noisy features, Euclidean distance is dominated by noise



KNN: Feature Weighting

- Scale each feature by its importance for classification

$$D(a,b) = \sqrt{\sum_k w_k (a_k - b_k)^2}$$

- Can use prior/domain knowledge
  - which features are more important
- Can learn the weights w_k using cross-validation

KNN: Computational Complexity

- Basic KNN algorithm stores all examples
- Suppose we have ***n*** examples each of dimension ***d***
- For each point to be classified
  - O(d) to compute distance to one example
  - O(nd) to compute distances to all examples
  - O(nk) time to find k closest examples
  - Total time: O(nk+nd)
- Very expensive for a large number of samples
- But we need a large number of samples for kNN to work well!

Reducing Complexity

- Various exact and approximate methods for reducing complexity
- Reduce dimensionality of the data
  - find projection to a lower dimensional space so that the distances between samples are approximately the same
  - PCA
  - Projection to a Random subspace
- Use smart data structures, like kd trees

KNN Summary

- Advantages
  - Can be applied to the data from any distribution
    - data does not have to be separable with a linear boundary
  - Simple and intuitive
  - Good classification with large number of samples
- Disadvantages
  - Choosing k may be tricky
  - Test stage is computationally expensive
    - No training stage, all the work is done during the test stage
    - This is actually the opposite of what we want. Usually we can afford training step to take a long time, but we want a fast test step
  - Need large number of samples for accuracy

References

- http://www.csd.uwo.ca/courses/CS9840a/Lecture2_knn.pdf
- http://classes.engr.oregonstate.edu/eecs/spring2012/cs534/notes/knn.pdf
- http://people.csail.mit.edu/dsontag/courses/ml12/slides/lecture10.pdf

Similarity Metrics

Matrix Dissimilarity

- Dissimilarity?

- $\begin{bmatrix} 3 & 5 \\ 6 & 9 \\ 11 & 21 \end{bmatrix}$

- How much input?

Matrix Dissimilarity

- Dissimilarity?

- $\begin{bmatrix} 3 & 5 \\ 6 & 9 \\ 11 & 21 \end{bmatrix}$

- How much input?
  - 3 $< x_1, x_2 >$ pairs

Matrix Dissimilarity

- Dissimilarity?
- $\begin{bmatrix} 3 & 5 \\ 6 & 9 \\ 11 & 21 \end{bmatrix}$
- Dissimilarity matrix size?

Matrix Dissimilarity

- Dissimilarity?

- $\begin{bmatrix} 3 & 5 \\ 6 & 9 \\ 11 & 21 \end{bmatrix}$

- Dissimilarity matrix size?
  - $3^2 = 9$

Matrix Dissimilarity

- Dissimilarity?

- $\begin{bmatrix} 3 & 5 \\ 6 & 9 \\ 11 & 21 \end{bmatrix}$

- $\begin{bmatrix} 0 & 0 & 0 \\ d(2,1) & 0 & 0 \\ d(3,1) & d(3,2) & 0 \end{bmatrix}$

Matrix Dissimilarity

- Dissimilarity?

- $\begin{bmatrix} 3 & 5 \\ 6 & 9 \\ 11 & 21 \end{bmatrix}$

- $\begin{bmatrix} 0 & 0 & 0 \\ \sqrt{(3-6)^2+(5-9)^2} & 0 & 0 \\ \sqrt{(3-11)^2+(5-21)^2} & \sqrt{(11-6)^2+(21-9)^2} & 0 \end{bmatrix}$

Matrix Dissimilarity

- Dissimilarity?

- $\begin{bmatrix} 3 & 5 \\ 6 & 9 \\ 11 & 21 \end{bmatrix}$

- Other distance functions?

Matrix Dissimilarity

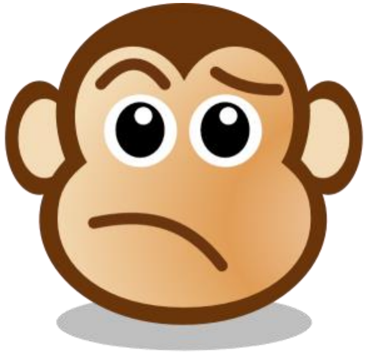
- Dissimilarity?
- $\begin{bmatrix} 3 & 5 \\ 6 & 9 \\ 11 & 21 \end{bmatrix}$
- $\begin{bmatrix} 0 & 0 & 0 \\ |3-6| + |5-9| & 0 & 0 \\ |3-11| + |5-21| & |11-6| + |21-9| & 0 \end{bmatrix}$

Nominal Attributes

- Dissimilarity between apple and orange?

Nominal Attributes

- Dissimilarity between apple and orange?



Nominal Attributes

- Student 1: likes jazz, eats pizza, roots for the cubs, wears socks
- Student 2: likes rock, eats pizza, roots for the cubs, goes barefoot
- $d(\text{Student 1, Student 2})$?

Nominal Attributes

- Student 1: likes jazz, eats pizza, roots for the cubs, wears socks
- Student 2: likes rock, eats pizza, roots for the cubs, goes barefoot
- $d(\text{Student 1, Student 2})$?
  - m: # of matches?

Nominal Attributes

- Student 1: likes jazz, eats pizza, roots for the cubs, wears socks
- Student 2: likes rock, eats pizza, roots for the cubs, goes barefoot
- $d(\text{Student 1, Student 2})$?
  - m: # of matches?
    - 2

Nominal Attributes

- Student 1: likes jazz, eats pizza, roots for the cubs, wears socks
- Student 2: likes rock, eats pizza, roots for the cubs, goes barefoot
- $d(\text{Student 1, Student 2})$?
  - m: # of matches?
    - 2
  - p: total # of variables?
    - 4

Nominal Attributes

- Student 1: likes jazz, eats pizza, roots for the cubs, wears socks
- Student 2: likes rock, eats pizza, roots for the cubs, goes barefoot
- $d(\text{Student 1, Student 2})$?
  - m: # of matches?
    - 2
  - p: total # of variables?
    - 4
  - $\frac{4-2}{4} = 0.5$

Nominal Attributes

- apple, banana, pear in binary?

Nominal Attributes

- apple, banana, pear in binary?
  - apple = 00
  - banana = 01
  - pear = 10

Binary Attributes

- symmetric?

Binary Attributes

- symmetric?
  - both outcomes equally important
  - Male/Female

Binary Attributes

- symmetric?
  - both outcomes equally important
  - Male/Female
- asymmetric?
  - not equally important
  - HIV positive/negative

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
  - How many positive for jack, negative for mary (r)?

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
  - How many positive for jack, negative for mary (r)?
    - 0…why not Jack=M and Mary=F

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
  - How many positive for jack, negative for mary (r)?
    - 0
  - How many negative for jack, positive for mary (s)?

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
  - How many positive for jack, negative for mary (r)?
    - 0
  - How many negative for jack, positive for mary (s)?
    - 1

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
    - How many positive for jack, negative for mary (r)?
        - 0
    - How many negative for jack, positive for mary (s)?
        - 1
    - How many positive for both? (q)

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
  - How many positive for jack, negative for mary (r)?
    - 0
  - How many negative for jack, positive for mary (s)?
    - 1
  - How many positive for both? (q)
    - 2

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)? r=0, s=1, q=2
  - $\frac{0+1}{2+0+1}$

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
    - How many positive for jack, negative for mary (r)?

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
  - How many positive for jack, negative for mary (r)?
    - 1

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
  - How many positive for jack, negative for mary (r)?
    - 1
  - How many negative for jack, positive for mary (s)?

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
  - How many positive for jack, negative for mary (r)?
    - 1
  - How many negative for jack, positive for mary (s)?
    - 0

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
  - How many positive for jack, negative for mary (r)?
    - 1
  - How many negative for jack, positive for mary (s)?
    - 0
  - How many positive for both? (q)

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
  - How many positive for jack, negative for mary (r)?
    - 1
  - How many negative for jack, positive for mary (s)?
    - 0
  - How many positive for both? (q)
    - 0

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)?
  - How many positive for jack, negative for mary (r)?
    - 1
  - How many negative for jack, positive for mary (s)?
    - 0
  - How many positive for both? (q)
    - 0
  - How many negative for jack, negative for mary (t)?
    - 0

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- d(jack, mary)? r=1, s=0, q=0,t=0
  - $\frac{1+0}{0+1+0+0}$

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- J(jack, mary)? r=0, s=1, q=2

Binary Attributes

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- Gender is symmetric, remaining asymmetric
- M, Y, P = 1 (positive), F,N (negative),N (no) = 0 (negative)
- J(jack, mary)? r=0, s=1, q=2
  - $\frac{2}{2+0+1}$

Ordinal Attributes

- Have order
  - freshman, sophomore, junior, senior

Ordinal Attributes

- Have order
  - freshman, sophomore, junior, senior
- freshman, sophomore, junior, senior mapped?

Ordinal Attributes

- Have order
  - freshman, sophomore, junior, senior
- freshman, sophomore, junior, senior mapped?
  - 1, 2, 3, 4

Ordinal Attributes

- Have order
  - freshman, sophomore, junior, senior
- freshman, sophomore, junior, senior mapped?
  - 1, 2, 3, 4
- Mapped onto [0,1]?

Ordinal Attributes

- Have order
  - freshman, sophomore, junior, senior
- freshman, sophomore, junior, senior mapped?
  - 1, 2, 3, 4
- Mapped onto [0,1]?
  - $\frac{1-1}{4-1}, \frac{2-1}{4-1}, \frac{3-1}{4-1}, \frac{4-1}{4-1}$

Ordinal Attributes

- Have order
  - freshman, sophomore, junior, senior
- freshman, sophomore, junior, senior mapped?
  - 1, 2, 3, 4
- Mapped onto [0,1]?
  - $\frac{1-1}{4-1}, \frac{2-1}{4-1}, \frac{3-1}{4-1}, \frac{4-1}{4-1}$
- Dissimilarity?

All Together

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

All Together

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- d(3,1)?

All Together

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

- d(3,1)?

All Together

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

Object Identifier	***test-1 (nominal)***	***test-2 (ordinal)***	***test-3 (numeric)***
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- d(3,1)?
  - $\frac{1(1)+1(0.5)+1(0.45)}{3}$

Cosine Similarity

- d1: I like to go the store
- d2: I like the cubs, go cubs go

Cosine Similarity

- d1: I like to go to the store
- d2: I like the cubs, go cubs go

Document	I	like	to	go	the	store	cubs
d1	1	1	2	1	1	1	0
d2	1	1	0	2	1	0	2

Cosine Similarity

- d1: I like to go to the store
- d2: I like the cubs, go cubs go

Document	I	like	to	go	the	store	cubs
d1	1	1	2	1	1	1	0
d2	1	1	0	2	1	0	2

- cos(d1, d2)?

Cosine Similarity

- d1: I like to go to the store
- d2: I like the cubs, go cubs go

Document	I	like	to	go	the	store	cubs
d1	1	1	2	1	1	1	0
d2	1	1	0	2	1	0	2

- cos(d1, d2)?
  - $\frac{1 \cdot 1 + 1 \cdot 1 + 2 \cdot 0 + 1 \cdot 2 + 1 \cdot 1 + 1 \cdot 0 + 0 \cdot 2}{\sqrt{1^2 + 1^2 + 2^2 + 1^2 + 1^2 + 1^2 + 0^2} \cdot \sqrt{1^2 + 1^2 + 0^2 + 2^2 + 1^2 + 0^2 + 2^2}}$

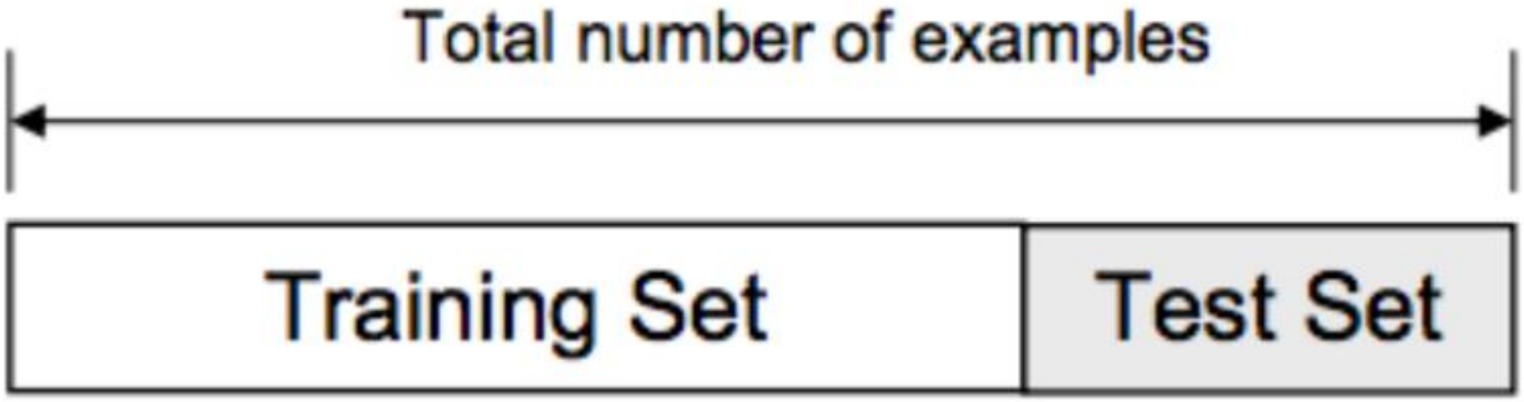
Classification Evaluation

Validation for Accuracy Estimation

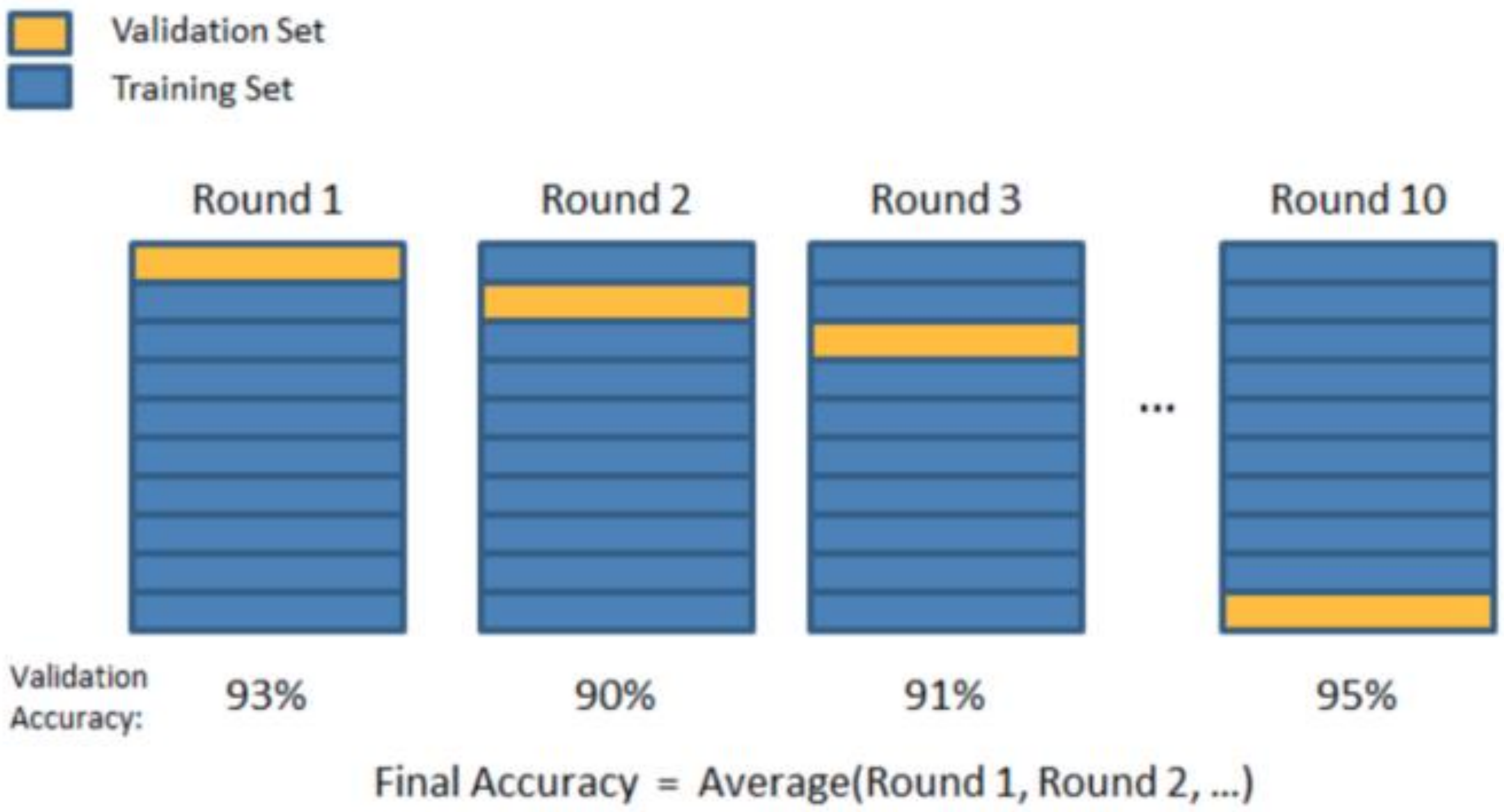
- Holdout method
  - (Randomly) partition data into two independent set
  - Train/test split

- Cross-validation (K-Fold CV)
  - Randomly partition the data into K mutually exclusive subsets
  - Iteratively use each subset as the test set and others as the training set
  - Leave-out-out (LOO): Let K be the number of instances

Holdout Method

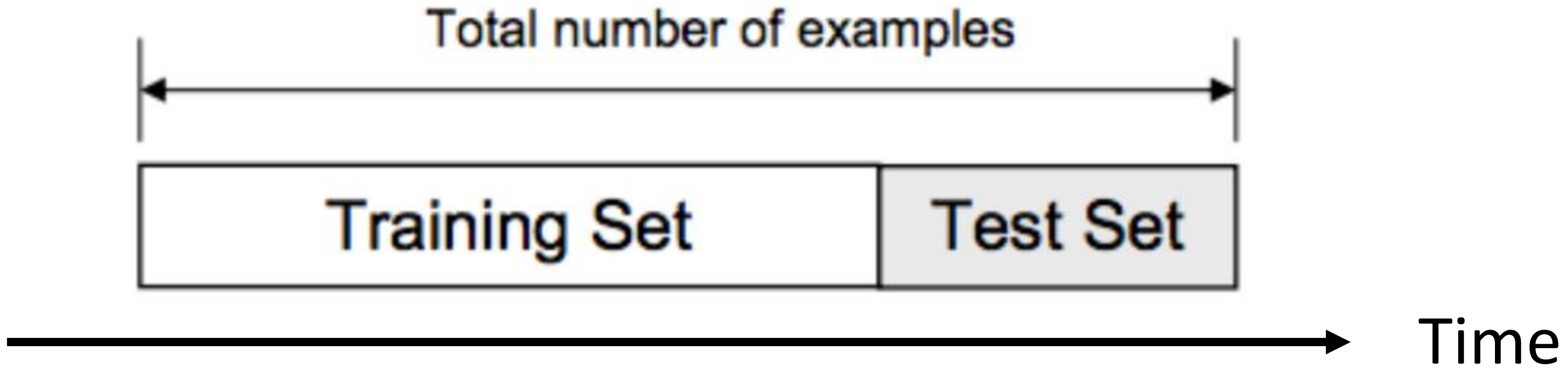


Cross Validation



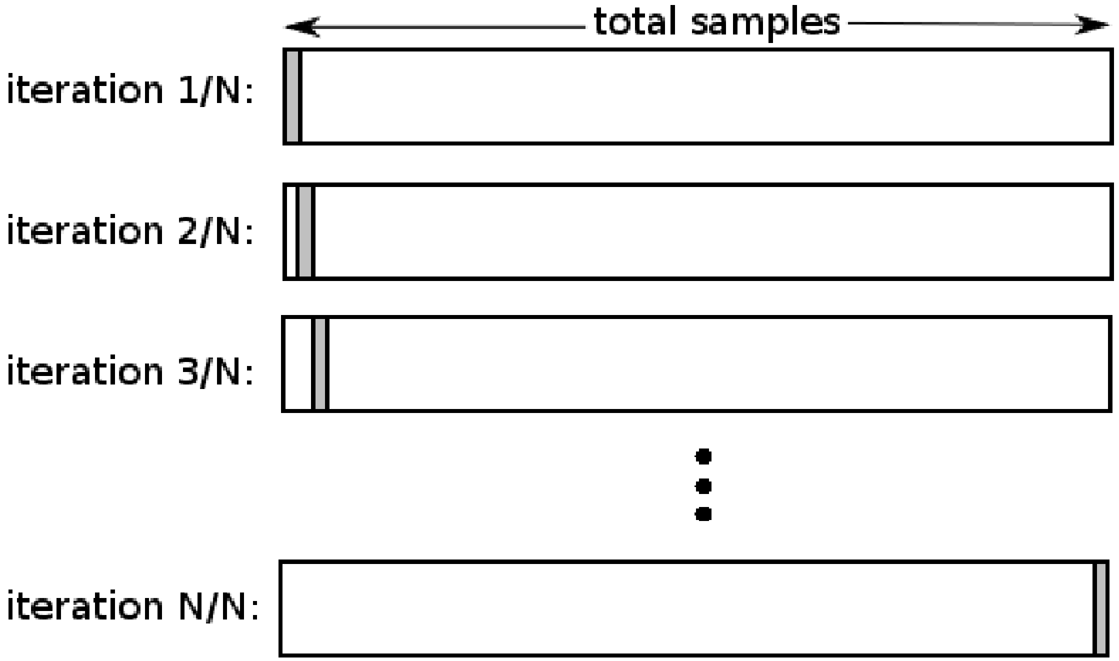
When to use holdout instead of CV?

- Cross validation is generally more comprehensive
- Some experimental settings may not allow to shuffle data.
    - Stock price prediction
    - Weather forecasting
    - …



Leave-One-Out (LOO) Cross Validation

- The most comprehensive evaluation approach
- Time-consuming if training the model is also time consuming.



Confusion Matrix

- Also called error matrix
- Usually used for binary classification
- Visualize information needed for performance evaluation
- Categorize predictions with correctness and classes

Actual class\Predicted class	C_1	$\neg C_1$
C_1	**True Positives (TP)**	**False Negatives (FN)**
$\neg C_1$	**False Positives (FP)**	**True Negatives (TN)**

Example of Confusion Matrix

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

Accuracy and Error Rate

- Accuracy = (TP + TN)/ALL
  - (6954+2588)/10000 = 0.9542
- Error rate = (FP + FN)/ALL = 1 - Accuracy
  - (412+46) / 10000 = 0.0458

A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
	P'	**N'**	**All**

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

Problem of Imbalance Data

- Some classes may be much rare
  - Fraud, HIV-positive
- High accuracy but unsatisfactory
  - 99% accuracy with all ~C predictions.
- Sensitivity: TP recognition rate
  - TP/P = 0/1 = 0%
- Specificity: TN recognition rate
  - TN/N = 99/99 = 99%

A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
	P’	**N’**	**All**

A\P	C	~C	
C	0	1	1
~C	0	99	99
	0	100	100

Precision and Recall

- Focus on a single class (usually positive class in binary classification)
- Precision: exactness, precision of positive predictions
  - TP / (TP + FP) = 6954 / (6954 + 412) = 0.9440
- Recall: completeness, recall for positive instances
  - TP / (TP + FN) = 6954 / (6954 + 46) = 0.9934

A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
	P’	**N’**	**All**

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

F-measures

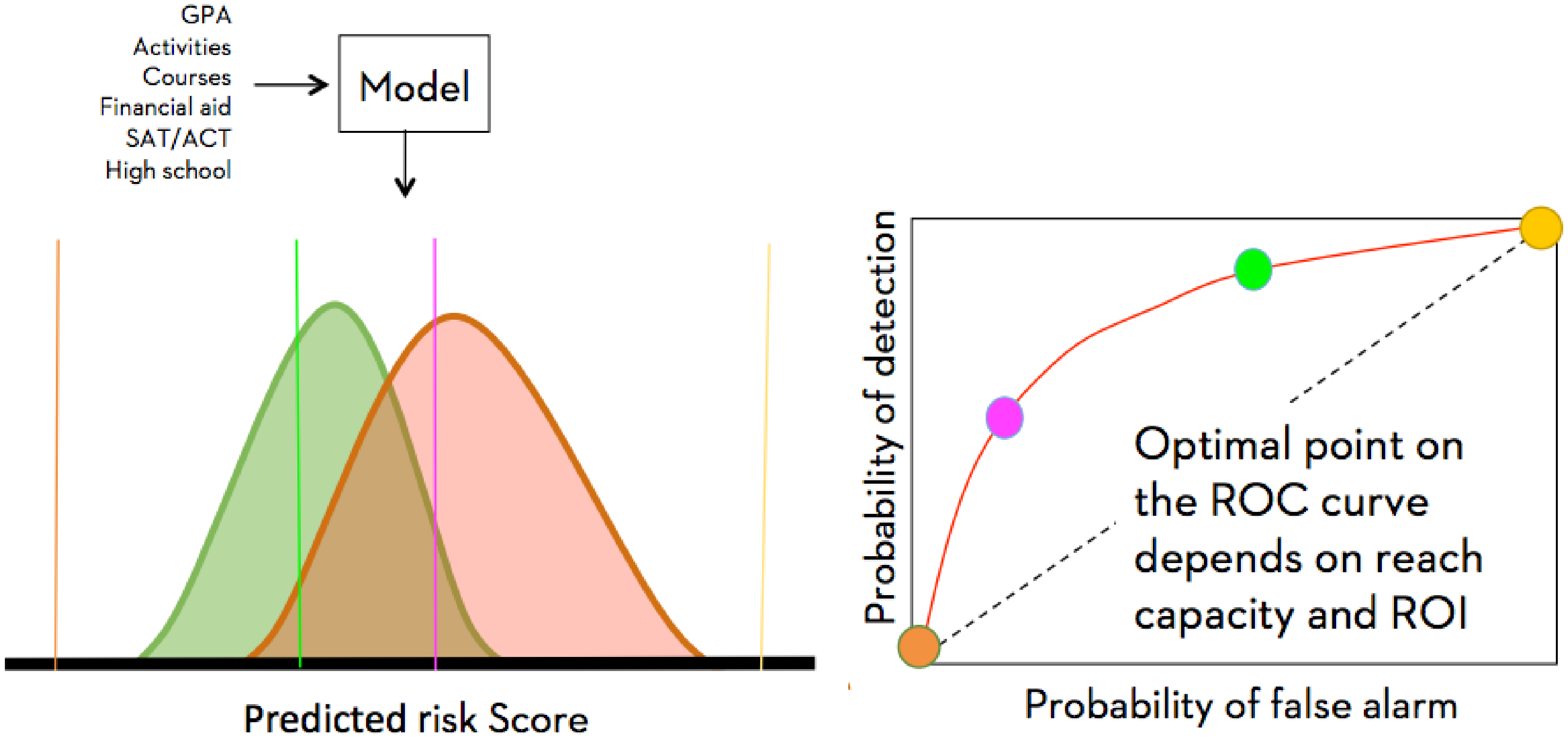
- Consider both precision (P) and recall (R)
- F_1 or F-score
  - $F = \frac{2 \times P \times R}{P+R} = \frac{2 \cdot 0.9440 \cdot 0.9934}{0.9440+0.9934} = 0.9681$
- F_β : weighted combination
  - $F = \frac{(1+\beta^2) \times P \times R}{\beta^2 \times P + R}$

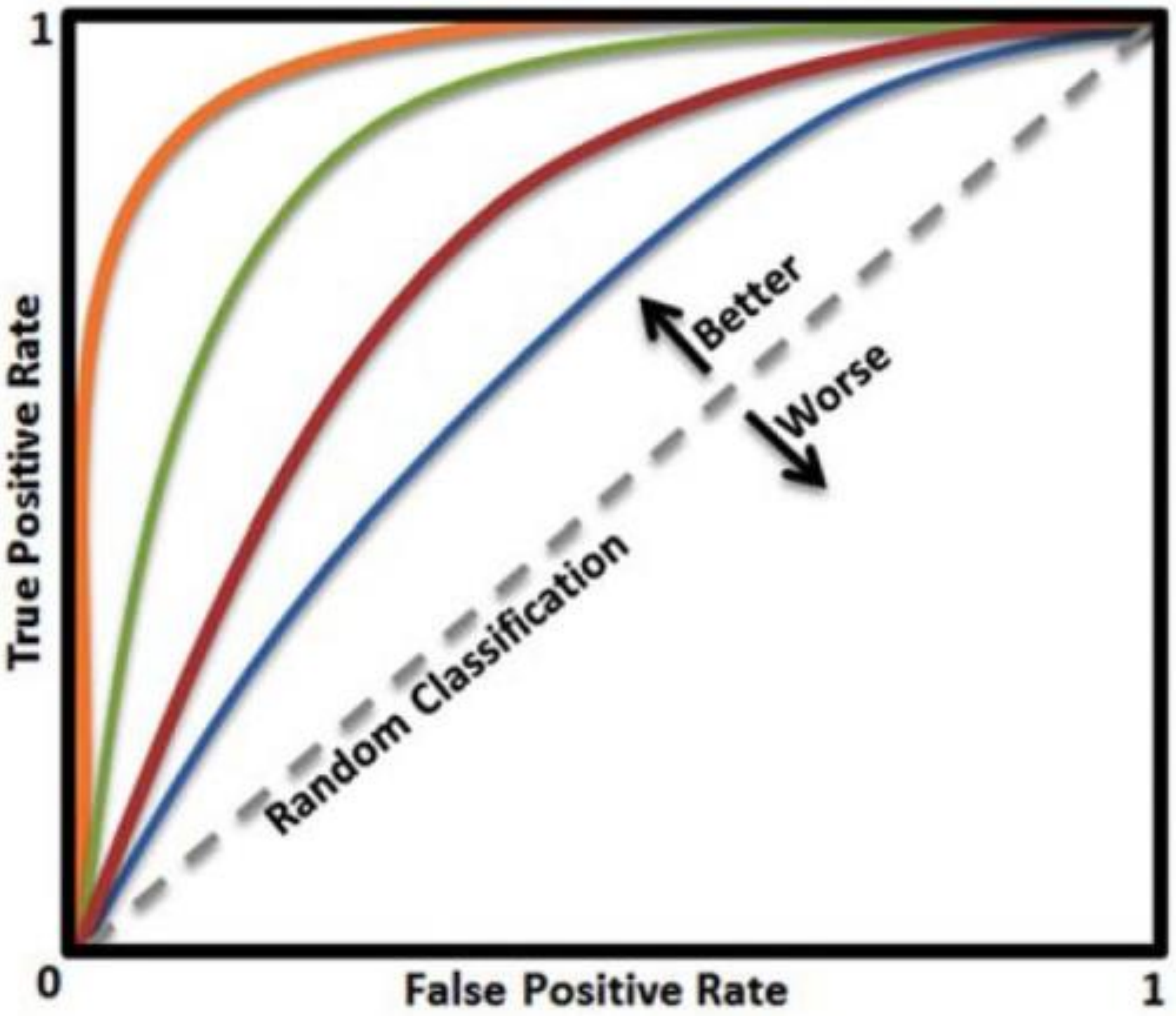
A\P	C	¬C	
C	**TP**	**FN**	**P**
¬C	**FP**	**TN**	**N**
	P'	**N'**	**All**

Actual class\Predicted class	buy_computer = yes	buy_computer = no	Total
buy_computer = yes	**6954**	**46**	7000
buy_computer = no	**412**	**2588**	3000
Total	7366	2634	10000

ROC Curves

- ROC (Receiver Operating Characteristics) curves
- Show the trade-off between true positive rate and false positive rate

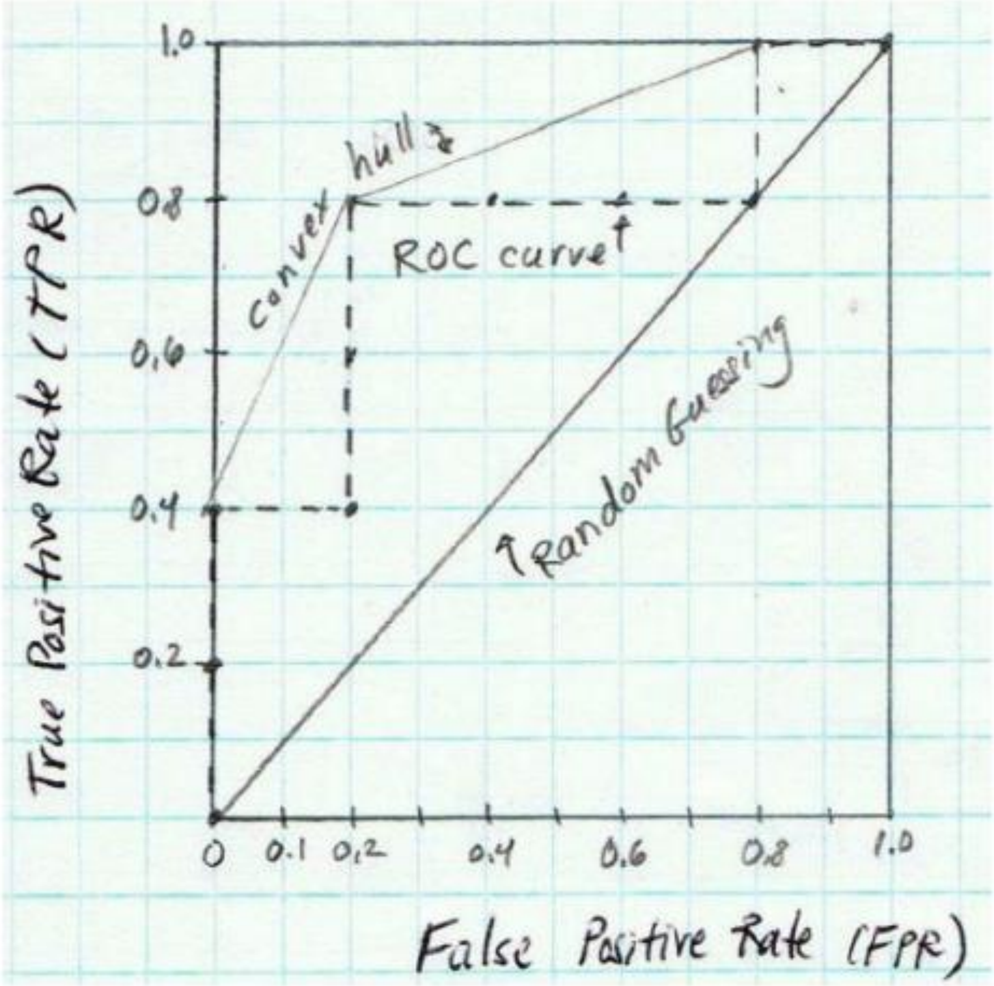




1
True Positive Rate
Better
Worse
Random Classification
0
False Positive Rate
1

Example of plotting an ROC curve

Tuple #	*Class*	*Prob.*	*TP*	*FP*	*TN*	*FN*	*TPR*	*FPR*
1	p	0.9	1	0	5	4	0.2	0
2	p	0.8	2	0	5	3	0.4	0
3	n	0.7	2	1	4	3	0.4	0.2
4	p	0.6	3	1	4	2	0.6	0.2
5	p	0.55	4	1	4	1	0.8	0.2
6	n	0.54	4	2	3	1	0.8	0.4
7	n	0.53	4	3	2	1	0.8	0.6
8	n	0.51	4	4	1	1	0.8	0.8
9	p	0.50	5	4	0	1	1.0	0.8
10	n	0.4	5	5	0	0	1.0	1.0



Accuracy vs. ROC Curves

- **Case 1:**

- You use an algorithm to identify students who are at risk of not continuing to the next term.
- Following the case study, 10% of students do not persist.
- You test your predictive model on the data and find that you made correct predictions 92% of the time.

Example from https://www.slideshare.net/KristenHunter/civitas-learning-understanding-roc-curves

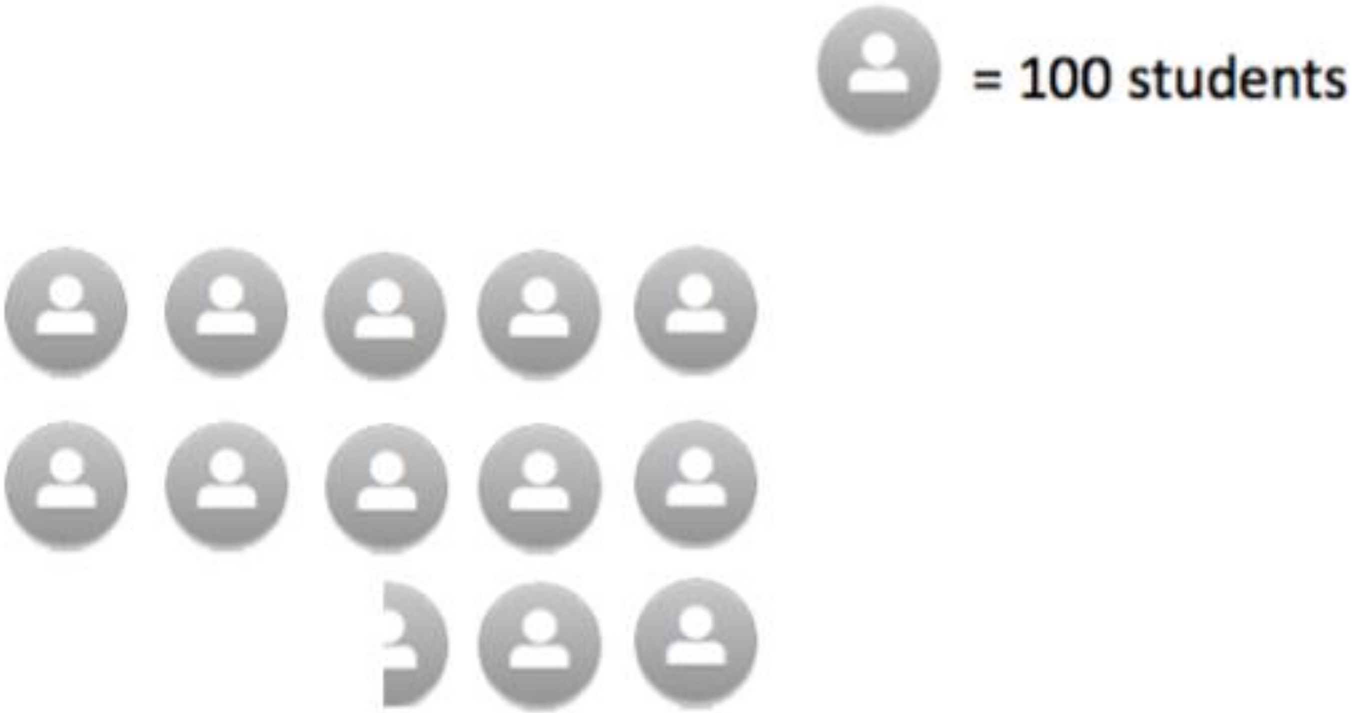
Accuracy vs. ROC Curves (Cont'd)

- A crackpot scientist tells you,
- "I could've gotten 90% accuracy just by predicting everyone will persist. After all the math, you gained only 2%?!"
- Don't give up yet!
- Your predictive model is still helpful.

Example from https://www.slideshare.net/KristenHunter/civitas-learning-understanding-roc-curves

Accuracy vs. ROC Curves (Cont’d)

- You have a team of advisors, and they have time to reach out to 1,250 students to suggest ways they can increase their likelihood of persisting



Example from https://www.slideshare.net/KristenHunter/civitas-learning-understanding-roc-curves

Accuracy vs. ROC Curves (Cont'd)

- **Without** the predictive model, you have to pick 1,250 students at random to assist. If 10% of them are expected to not persist, only 125 students would be likely to benefit from the intervention.

Example from https://www.slideshare.net/KristenHunter/civitas-learning-understanding-roc-curves

Accuracy vs. ROC Curves (Cont'd)

- **With** the predictive model, you can choose the 1,250 students by ordering them by the highest predicted risk score.
- The test case reveals 600 of these students are at risk and would be most likely to benefit from the right intervention at the right time.

Example from https://www.slideshare.net/KristenHunter/civitas-learning-understanding-roc-curves

The ROC Curve Trade-off

Students most likely to benefit from an intervention

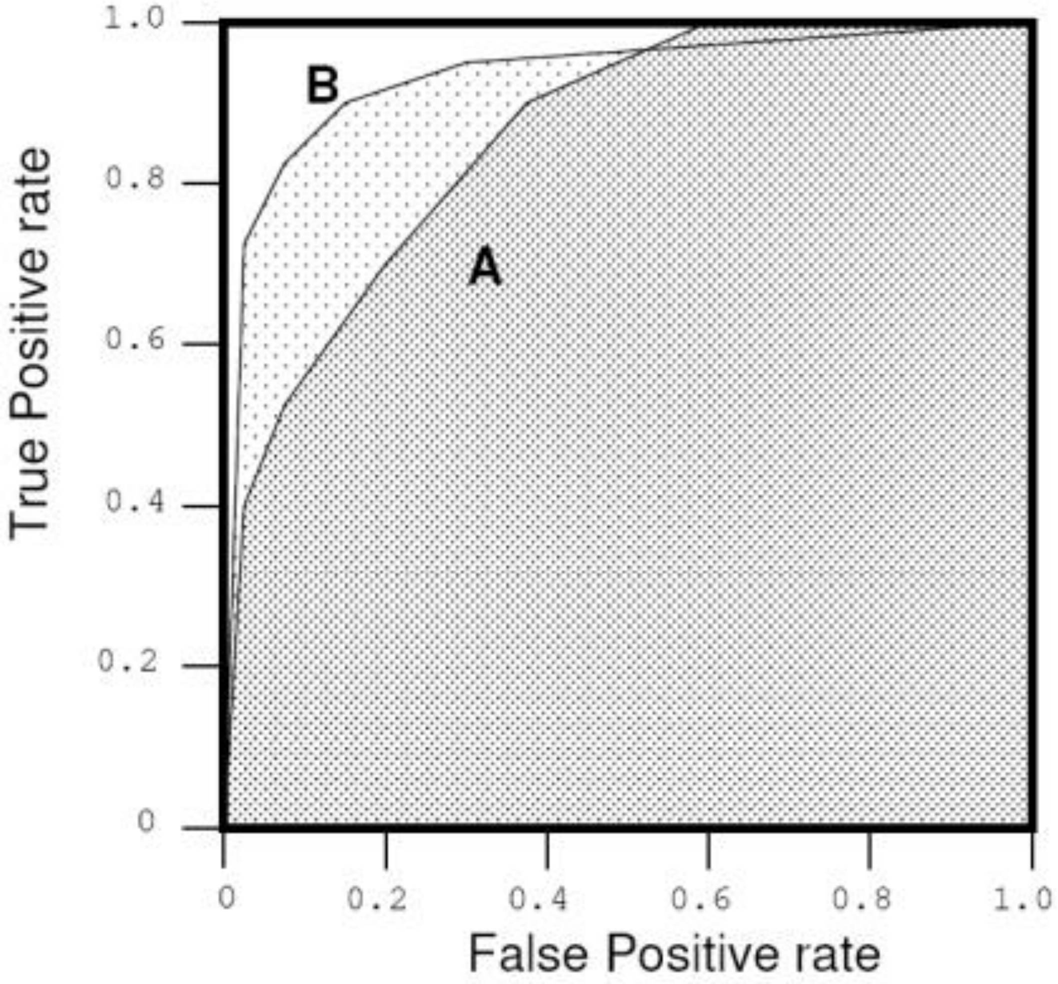
WITHOUT PREDICTIVE MODEL	WITH PREDICTIVE MODEL
125 students	**600 students**

~5x improvement

Example from https://www.slideshare.net/KristenHunter/civitas-learning-understanding-roc-curves

Area under a ROC Curve (AUC)

- A standard evaluation metric with a ROC curve
- Can be computed while constructing RUC curves
- From 0 (worst) to 1 (best)
- Equivalent to pairwise accuracy



Reference

- http://web.cs.ucla.edu/~yzsun/classes/2017Fall_CS145/Slides/08Evaluation_Classification.pdf
- https://www.slideshare.net/KristenHunter/civitas-learning-understanding-roc-curves
- https://web.uvic.ca/~maryam/DMSpring94/Slides/9_roc.pdf