CS 66: Machine Learning

Prof. Sara Mathieson

Spring 2019

SWARTHMORE

Outline for April 22

- Lab 5 Analysis Questions
- Midterm 2 Review
- Practice Problems
- Questions for Wednesday

- Final Project Proposals: all feedback and repos finished
- Wednesday in class "office hours" **(submit a question today!)**
- Fri: Guest lecture by Prof. Matt Zucker
- **Office hours today: 12:30-2pm & 3-4pm**
- I will not be on campus tomorrow, so make sure to come to office hours today **(and also post on Piazza!)**

Outline for April 22

- Lab 5 Analysis Questions
- Midterm 2 Review
- Practice Problems
- Questions for Wednesday

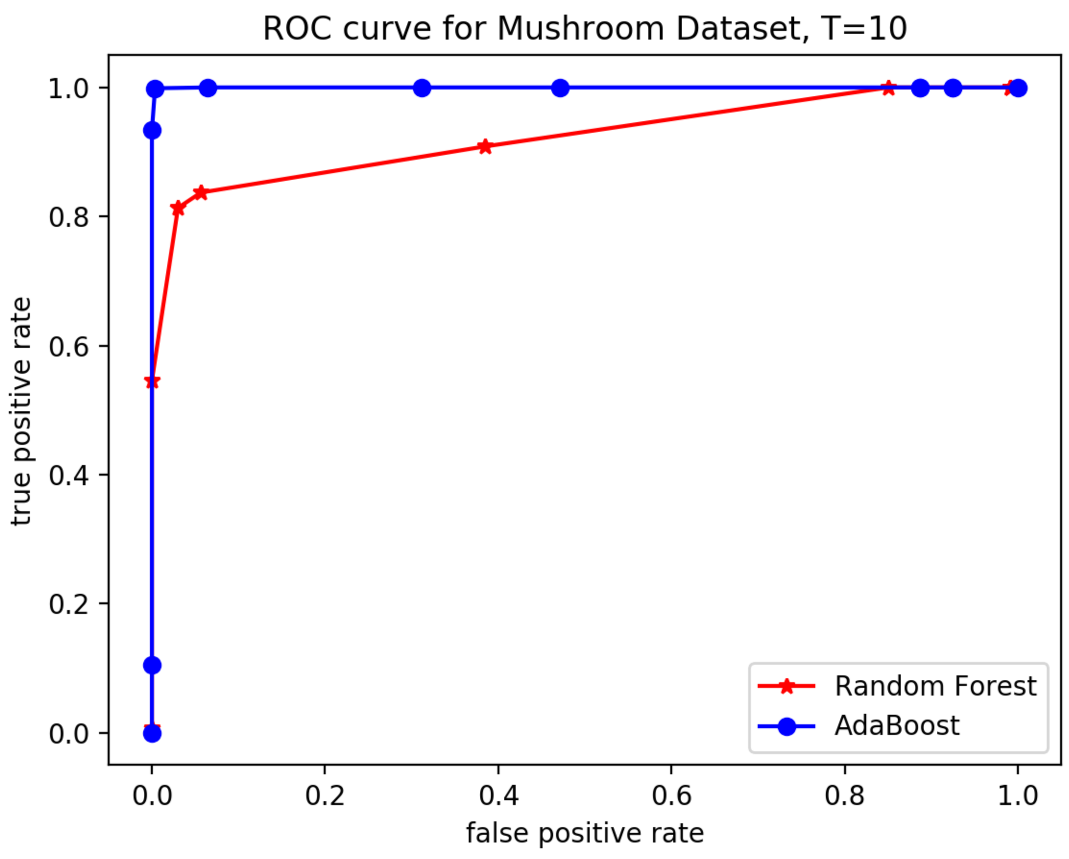
Lab 5: Runtimes

- n=number of examples, p=number of features, T=number of classifiers
- Random Forests: O(sqrt(p)nT)
- AdaBoost: O(pnT)
- Random forests is better even given this analysis, but it is also very parallelizable! AdaBoost is not

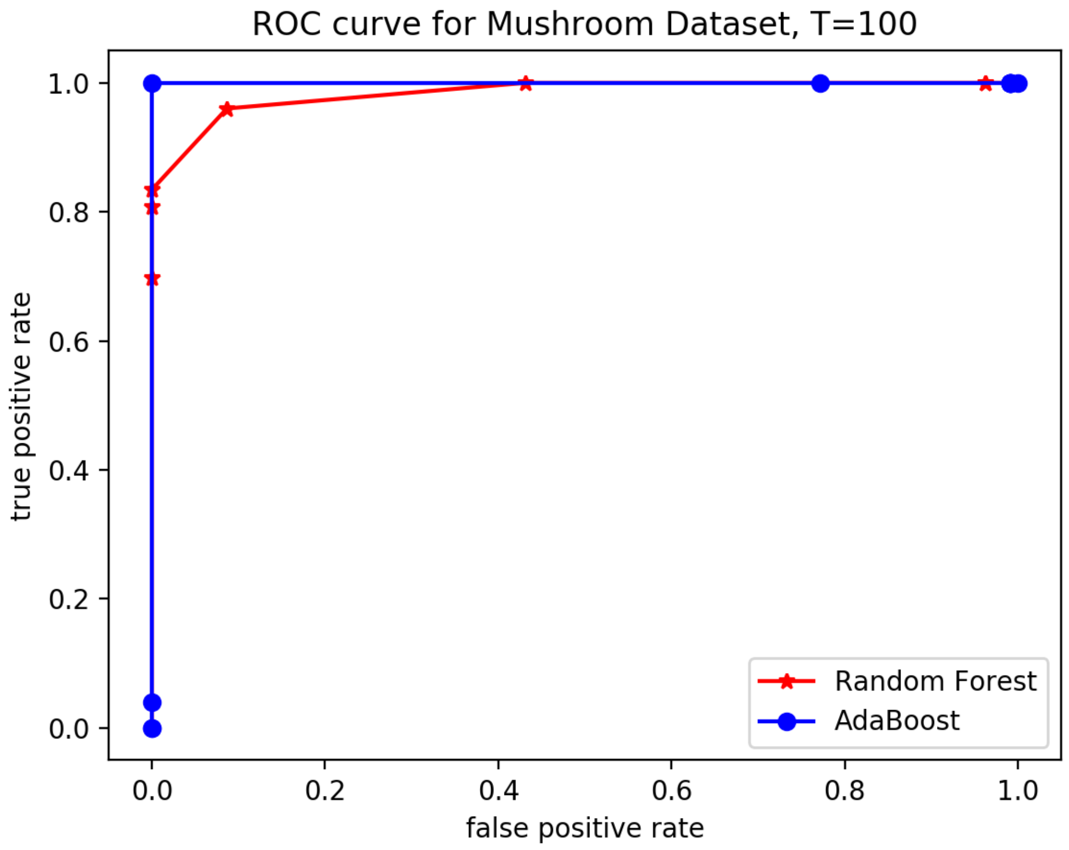
Ensemble Methods: overfitting

- Ensemble methods are very robust to overfitting!
- If all classifiers in the ensemble are “weak”, then nothing about the overall model is fit to the “noise” in the data
- Very powerful idea, and one reason why ensemble methods are still widely used

Lab 5: ROC curve (T=10)



Lab 5: ROC curve (T=100)



Outline for April 22

- Lab 5 Analysis Questions
- Midterm 2 Review
- Practice Problems
- Questions for Wednesday

Naïve Bayes

* pickup notecard & handout
* write $\boxed{\text{Question}}$ on card for Wednesday
* can use double sided cheat sheet for exam!

Naive Bayes

$$\underbrace{p(y=k|\vec{x})}_{\text{posterior}} = \frac{\overbrace{p(y=k)}^{\text{prior}}\ \overbrace{p(\vec{x}|y=k)}^{\text{likelih}}}{\underbrace{p(\vec{x})}_{\text{evidence}}}$$

$k = 1, 2, \ldots K$

# of classes

$$P(A,B|C)$$

$$= P(A|B,C)\,P(B|$$

Naive Bayes

$$\underbrace{p(y=k|\vec{x})}_{\text{posterior}} = \frac{\overbrace{p(y=k)}^{\text{prior}}\,\overbrace{p(\vec{x}|y=k)}^{\text{likelihood}}}{\underbrace{p(\vec{x})}_{\text{evidence}}}$$

$k = 1, 2, \ldots K$

# of classes

$$P(A,B|C)$$

$$= P(A|B,C)\,P(B|C)$$

$$p(\vec{x}|y=k)$$

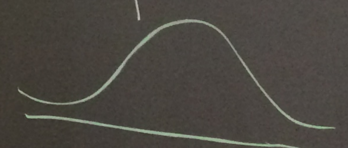
$$= p(\overbrace{x_1}^{A}, \overbrace{x_2 \cdots x_p}^{B} \,|\, \overbrace{y=k}^{C})$$

$$= p(x_1|\cancel{x_2 \cdots x_p}, y=k)\,p(x_2 \cdots x_p|y=k)$$

$$\approx \prod_{j=1}^{p} p(x_j|y=k)$$

NB assumption.

if cont. fit a Gaussian



Naïve Bayes Assumption

- Feature j is independent of all other features, given the class label

X_1	$\cdots$	X_5	$\cdots$	X_p	Y
C	$\cdots$	A	$\cdots$	B	4

estimat

train

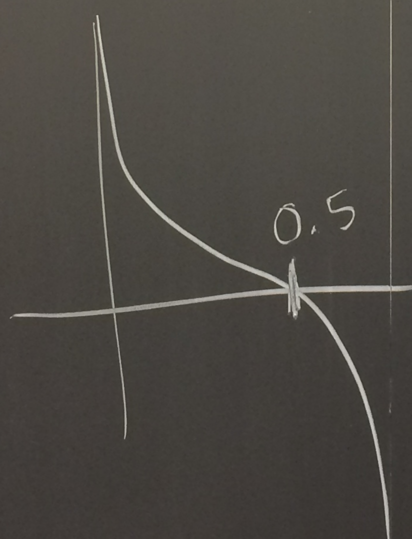
no

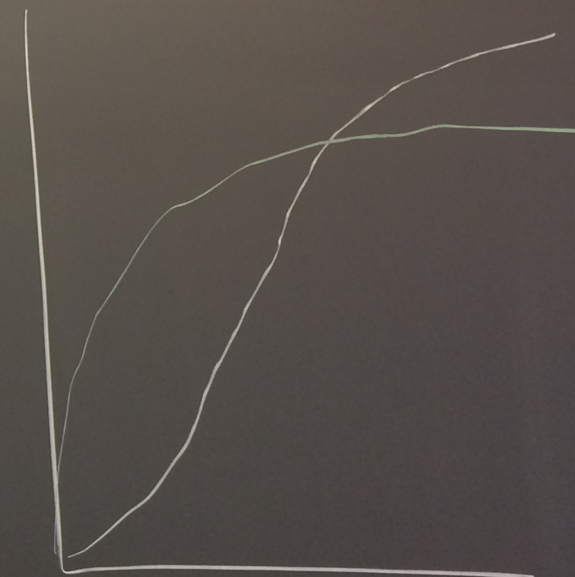
estimate $p(x_j \mid y=k) = \dfrac{P(x_j, y=k)}{P(y=k)}$

$$\approx \frac{N_{j,v,k}+1}{N_k + |f_j|}$$

train x_p & y

no $B \quad 4$





Evaluation Metrics

Recap Precision and Recall

- Precision: of all the “flagged” examples, which ones are actually relevant (i.e. positive)?

(Purity)

- Recall: of all the relevant results, which ones did I actually return?

(Completeness)

Modified from Ameet Soni

Recap Confusion Matrices

		Predicted class	
True class	Negative	Positive	
Negative	True negative (TN)	False positive (FP) “false alarm”	N (total number of true negatives)
Positive	False negative (FN) “miss”	True positive (TP)	P (total number of true positives)
	N* (what we said was negative)	P* (what we said was positive “flagged”)	

Modified from Jessica Wu

Recap Confusion Matrices

Predicted class

		Negative	Positive	
True class	Negative	True negative (TN)	False positive (FP) “false alarm”	N
	Positive	False negative (FN) “miss”	True positive (TP)	P
		N*	P*	

Modified from Jessica Wu

Recap Confusion Matrices

Predicted class

True class

	Negative	Positive	
Negative	True negative (TN)	False positive (FP) “false alarm”	N
Positive	False negative (FN) “miss”	True positive (TP)	P
	N*	P*	

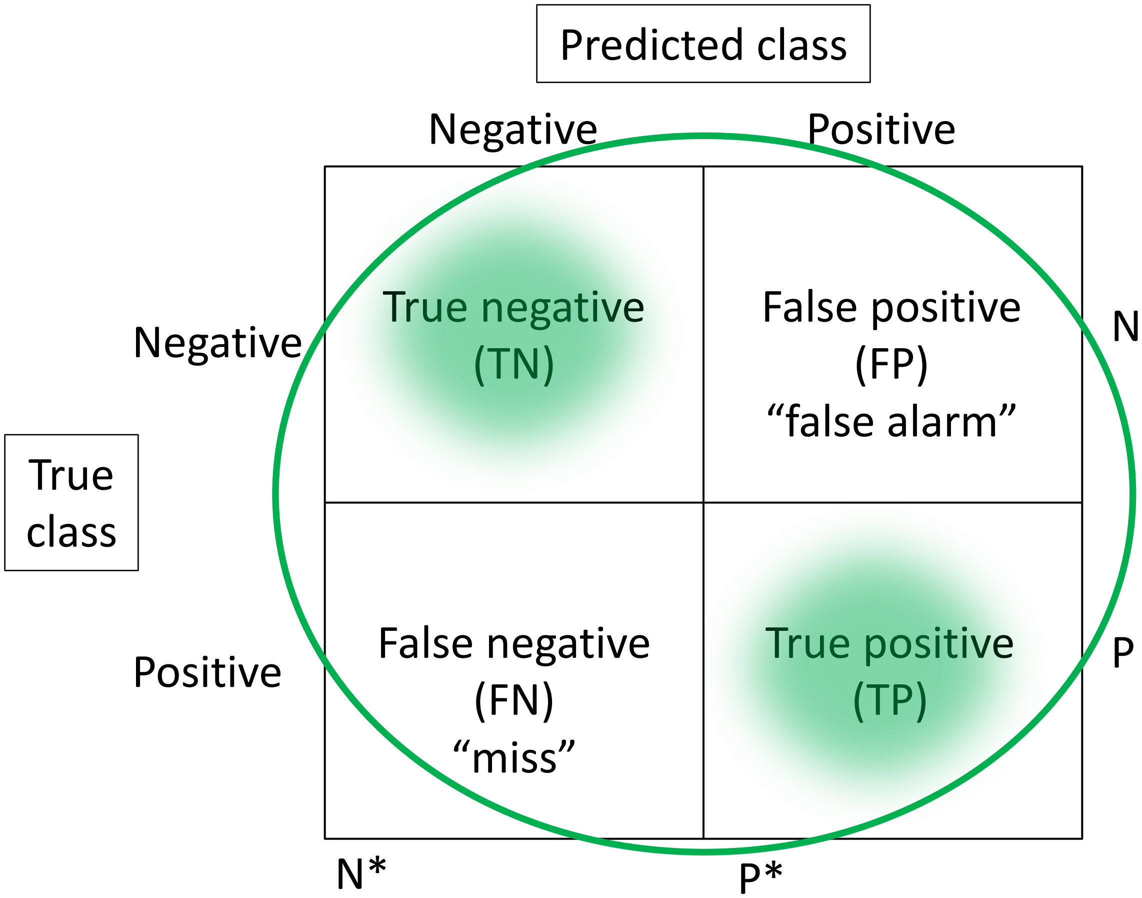
Error:

$(FN+FP)/(TN+FP+FN+TP)$

$= (FN+FP)/(N+P)$

Modified from Jessica Wu

Recap Confusion Matrices



Accuracy = 1-Error:

(TN+TP)/(TN+FP+FN+TP)

= (TN+TP)/(N+P)

Recap Confusion Matrices

		Predicted class		
		Negative	Positive	
True class	Negative	True negative (TN)	False positive (FP) “false alarm”	N
	Positive	False negative (FN) “miss”	True positive (TP)	P
		N*	P*	

Precision:

TP/(FP+TP) = TP/P*

Modified from Jessica Wu

Recap Confusion Matrices

Predicted class

True class	Negative	Positive	
Negative	True negative (TN)	False positive (FP) “false alarm”	N
Positive	False negative (FN) “miss”	True positive (TP)	P
	N*	P*	

Recall
(True Positive Rate):

TP/(FN+TP) = TP/P

Modified from Jessica Wu

Recap Confusion Matrices

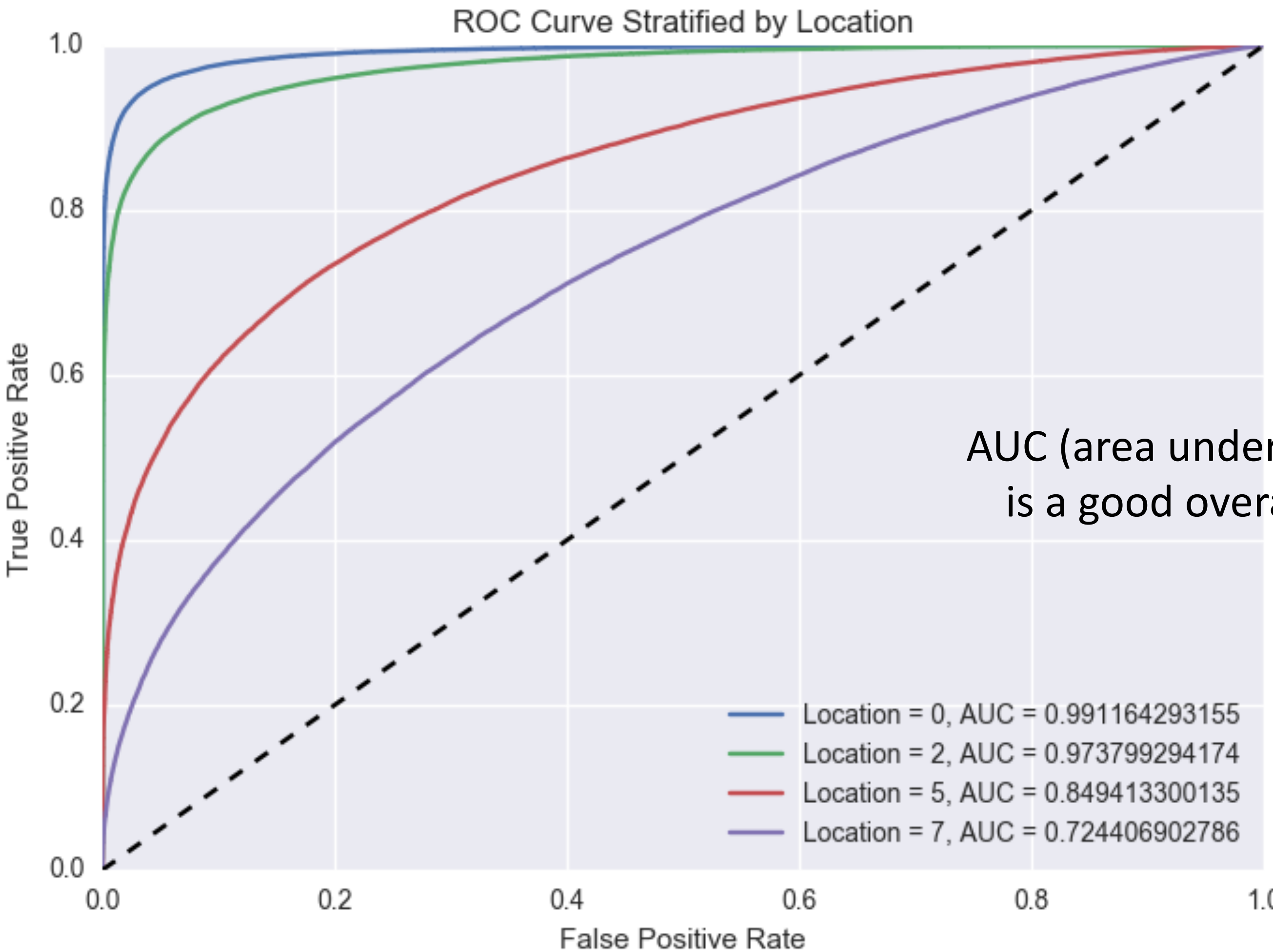
		Predicted class		
		Negative	Positive	
True class	Negative	True negative (TN)	False positive (FP) “false alarm”	N
	Positive	False negative (FN) “miss”	True positive (TP)	P
		N*	P*	

False Positive Rate:

FP/(TN+FP) = FP/N

Modified from Jessica Wu

ROC curve example: comparing methods



AUC (area under the curve) is a good overall metric

Example of a ROC curve from my research
Chan, Perrone, Spence, Jenkins, Mathieson, Song

Cross Validation

- Allows us to choose best hyper-parameters
- Allows us to return multiple independent accuracy results
- We can use this distribution of accuracy numbers in statistical frameworks (find mean/variance, compare with other methods, etc)

Ensemble Methods

Learning Theory

Let H be the hypothesis space

Three sources of limitations for traditional classifiers:

- Statistical - H is too large relative to size of data
  - Many hypotheses can fit the data by chance
- Computational - H is too large to completely search for "best" model
- Representational - H is not expressive enough

Learning Theory

- Statistical: Average of unstable models (high variance) has more stability
- Computational: searching from multiple starting points is better approximation than one starting point
- Representational: sum of many models can represent more hypotheses than an individual model

Ensembles can address all 3!

Learning Theory

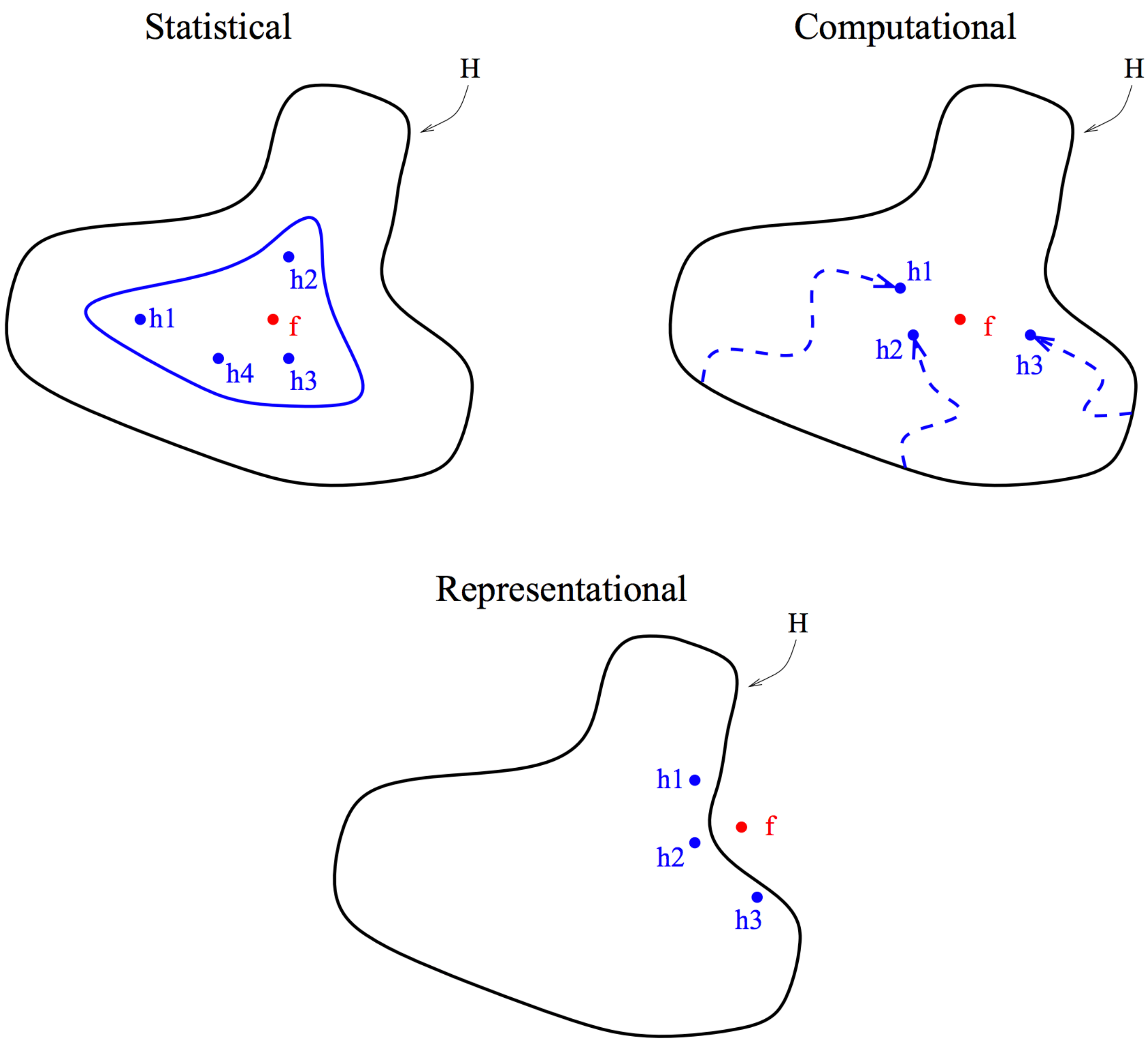


Figure from Tom Dietterich

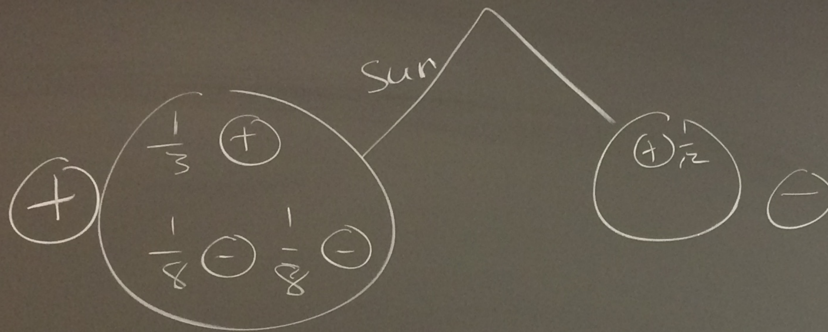
Outline for April 22

- Lab 5 Analysis Questions
- Midterm 2 Review
- Practice Problems
- Questions for Wednesday

=k)

t. fit

Gaussian



$$P(+) = \frac{\frac{1}{3}}{\frac{1}{3}+\frac{1}{8}+\frac{1}{8}} = \frac{12}{21} = \frac{4}{7}$$

$$\frac{4}{7} \geq 0.5$$

$$\varepsilon_t = \frac{1}{8} + \frac{1}{8} + \frac{1}{12} = \frac{1}{3}$$

Handout 19, Question 1, parts (a) and (b)

Handout 19, Question 2

$r = \frac{1}{3}$, $T = 5$ $T = 4$

R = # votes for <u>wrong</u> class

$R = 0, 1, 2, \underbrace{3, 4, 5}_{\text{pt classified incorrectly overall}}$

$$P(R = k) = \binom{T}{k} \underbrace{r^k}_{\text{wrong}} \underbrace{(1-r)^{T-k}}_{\text{right}}$$

$$P(R = 5) = \binom{5}{5}\left(\frac{1}{3}\right)^5$$

$$P(R = 4) = \binom{5}{4}\left(\frac{1}{3}\right)^4\left(\frac{2}{3}\right)$$

$$P(R = 3) = \binom{5}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^2$$

$0! = 1$

$\binom{5}{0} = 1$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

$$\sum_{k=\lceil \frac{T+1}{2} \rceil}^{T} P(R = k) \approx 0.21$$

Overall prob of being wrong.

$\frac{4+1}{2} = \lceil 2.5 \rceil \to 3$