MACHINE LEARNING OVERVIEW

MACHINE LEARNING PROBLEMS

	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

CLASSIFIER OVERVIEW

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

MANY CLASSIFIERS TO CHOOSE FROM

K-nearest neighbor

Support Vector Machines

Decision Trees

Random Forrest

(Gradient) Boosted Decision Trees

Logistic Regression

Naïve Bayes

Bayesian network

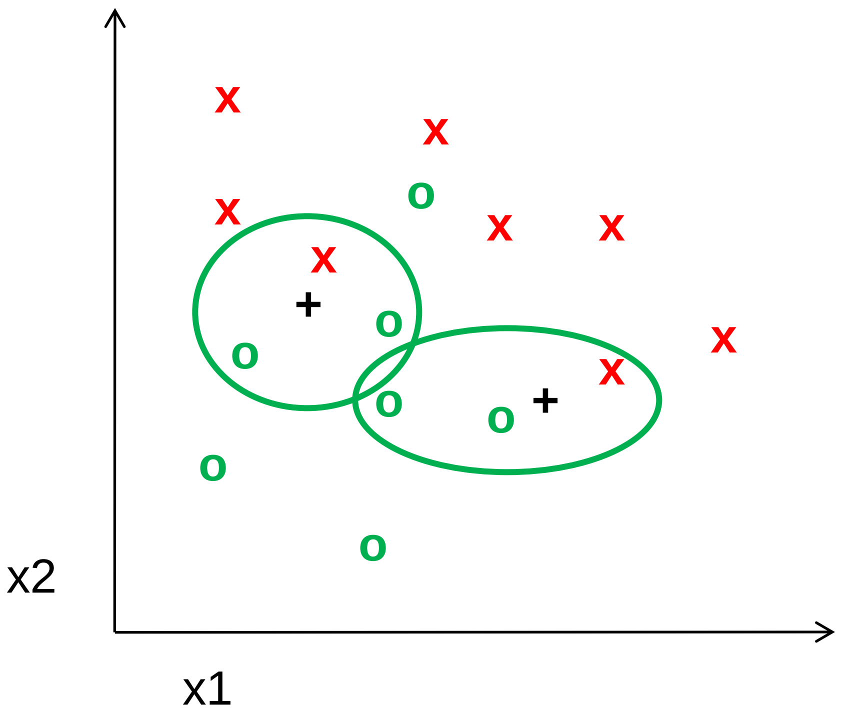
RBMs

....

Which is the best one?

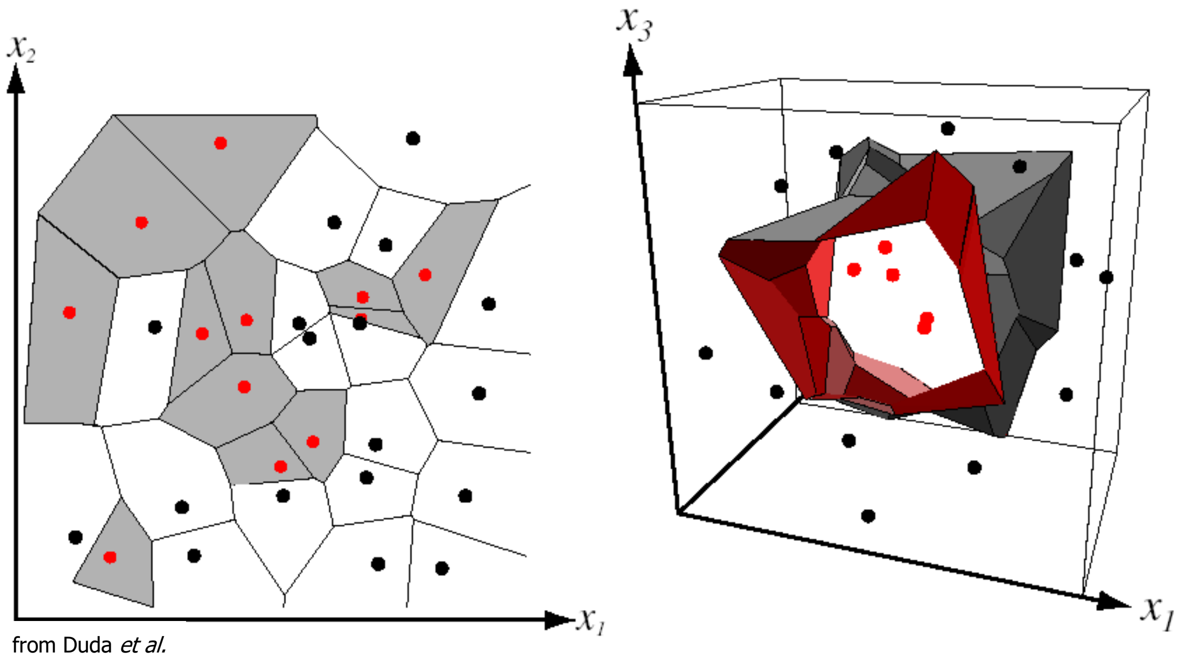
Slide credit: D. Hoiem

3-NEAREST NEIGHBOR



DECISION BOUNDARIES KNN

Assign label of nearest training data point to each test data point



Voronoi partitioning of feature space for two-category 2D and 3D data

Source: D. Lowe

MANY CLASSIFIERS TO CHOOSE FROM

K-nearest neighbor

Support Vector Machines

Decision Trees

Random Forrest

(Gradient) Boosted Decision Trees

Logistic Regression

Naïve Bayes

Bayesian network

RBMs

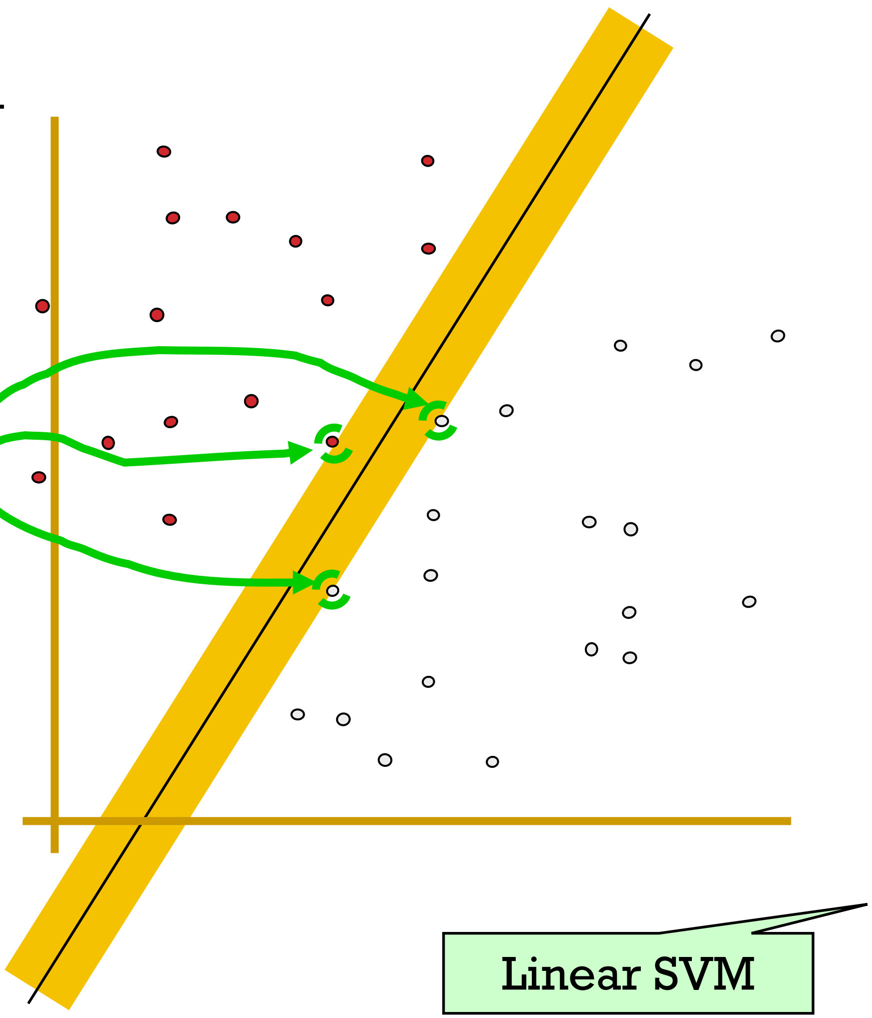
….

Which is the best one?

Slide credit: D. Hoiem

MAXIMUM MARGIN

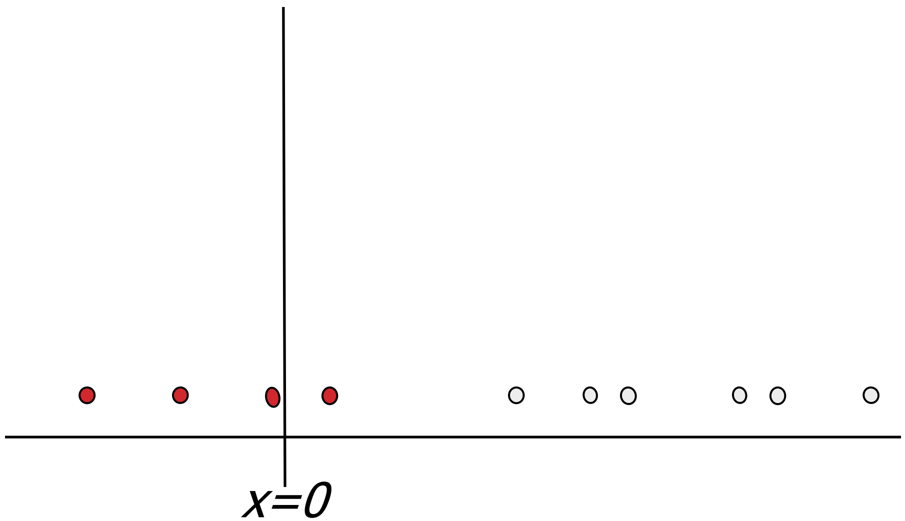
● denotes +1
○ denotes -1



Support Vectors are those datapoints that the margin pushes up against

The maximum margin linear classifier is the linear classifier with the, um, maximum margin. This is the simplest kind of SVM (Called an LSVM)

SUPPOSE WE'RE IN 1-DIMENSION

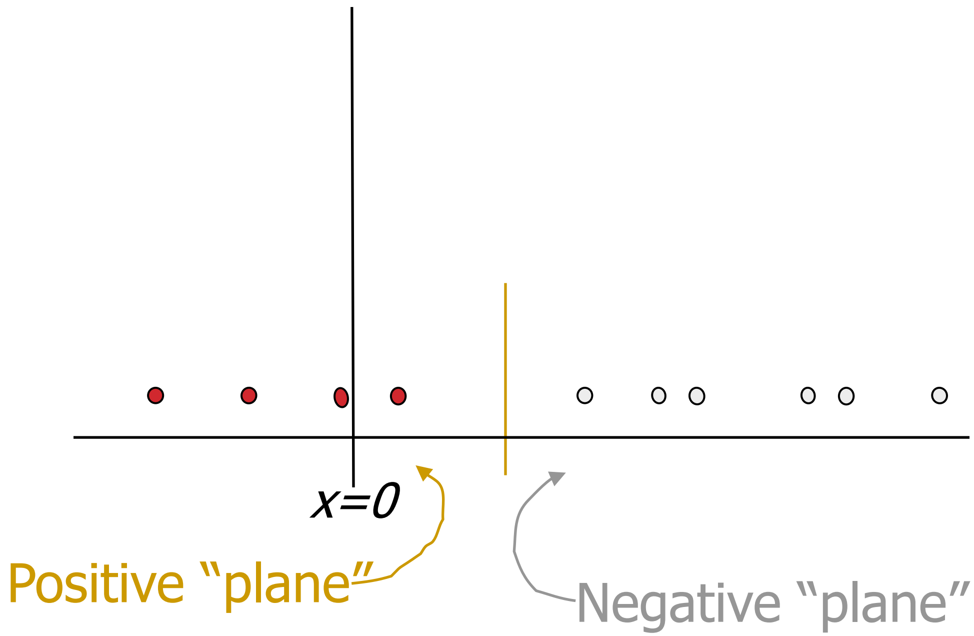


What would SVMs do with this data?

Slides from Andrew W. Moore

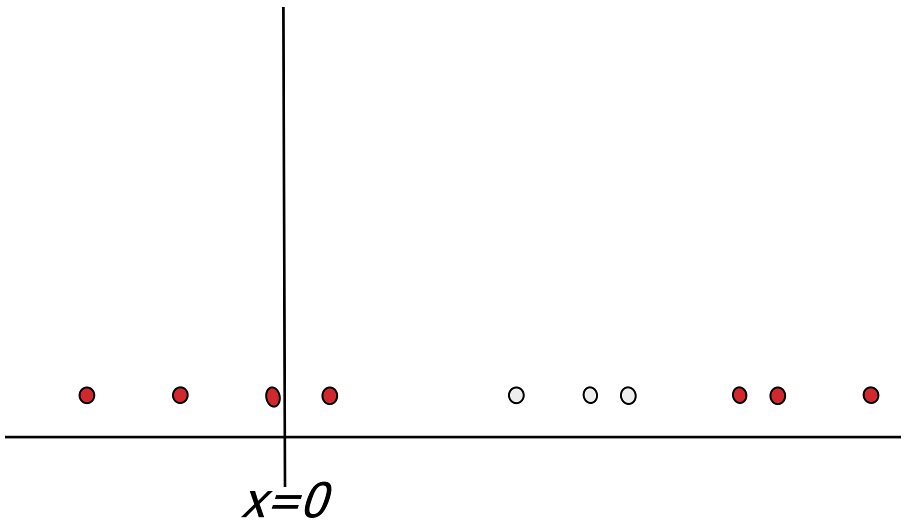
SUPPOSE WE'RE IN 1-DIMENSION

Not a big surprise



Slides from Andrew W. Moore

HARDER 1-DIMENSIONAL DATASET

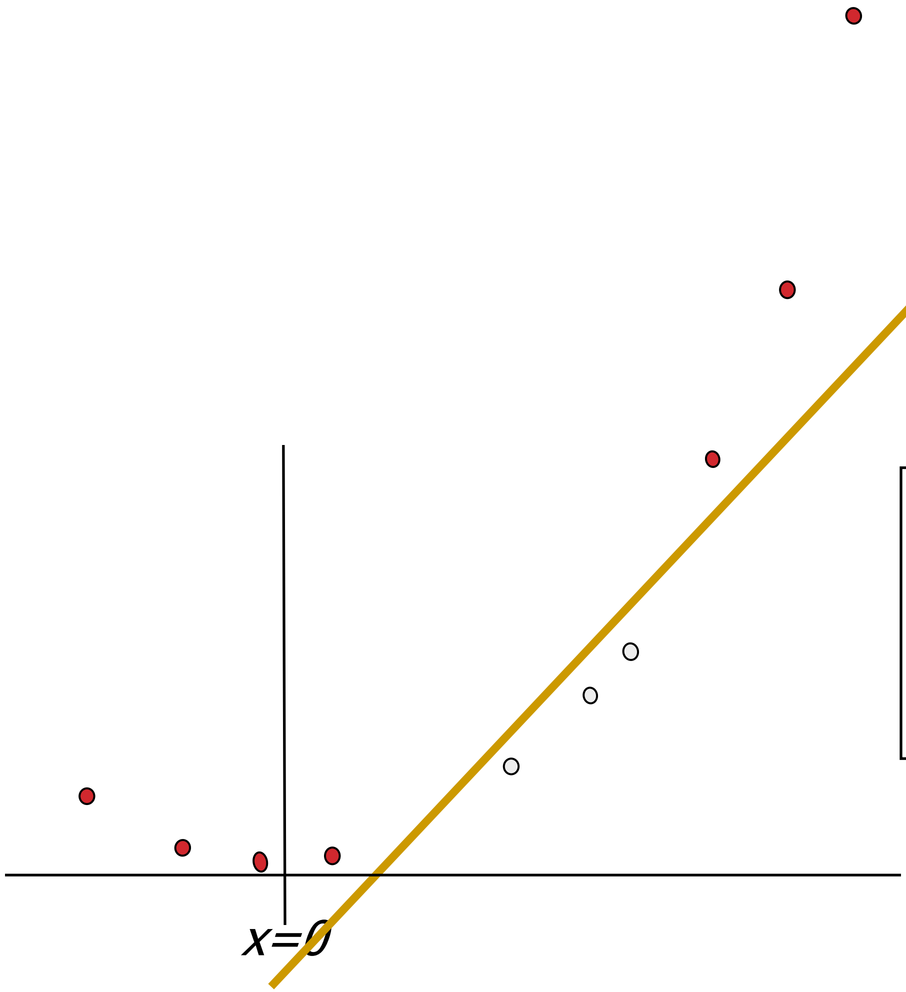


That's wiped the smirk off SVM's face.

What can be done about this?

Slides from Andrew W. Moore

HARDER 1-DIMENSIONAL DATASET

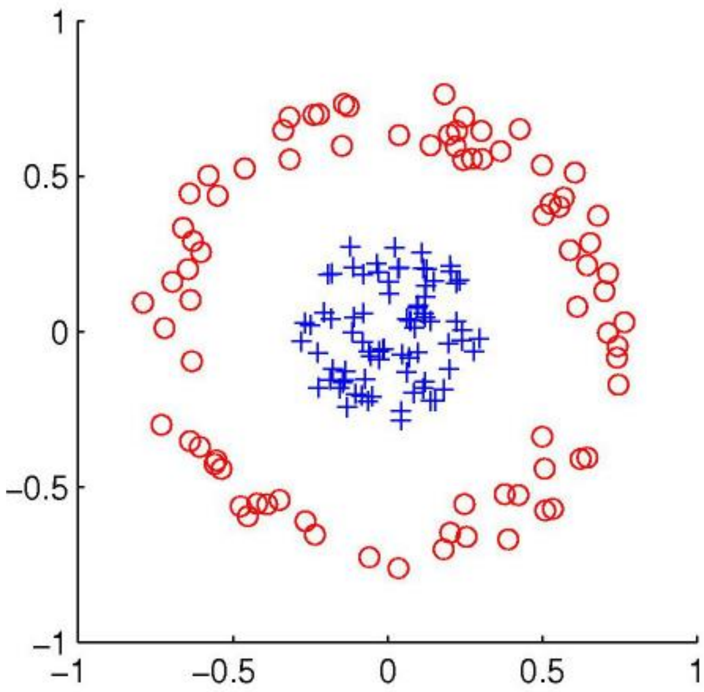


Permitting non-linear basis functions

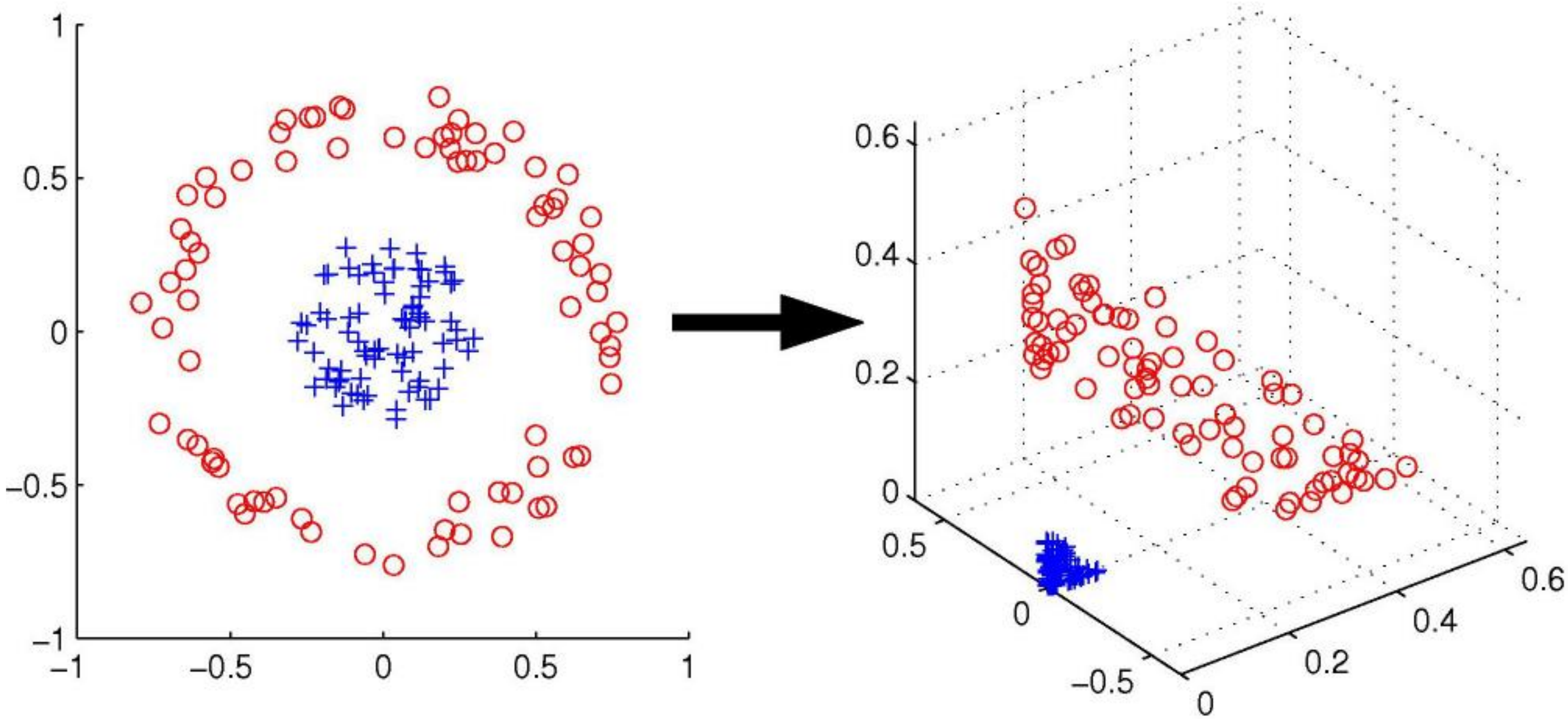
$$\mathbf{z}_k = (x_k, x_k^2)$$

Slides from Andrew W. Moore

THE KERNEL TRICK



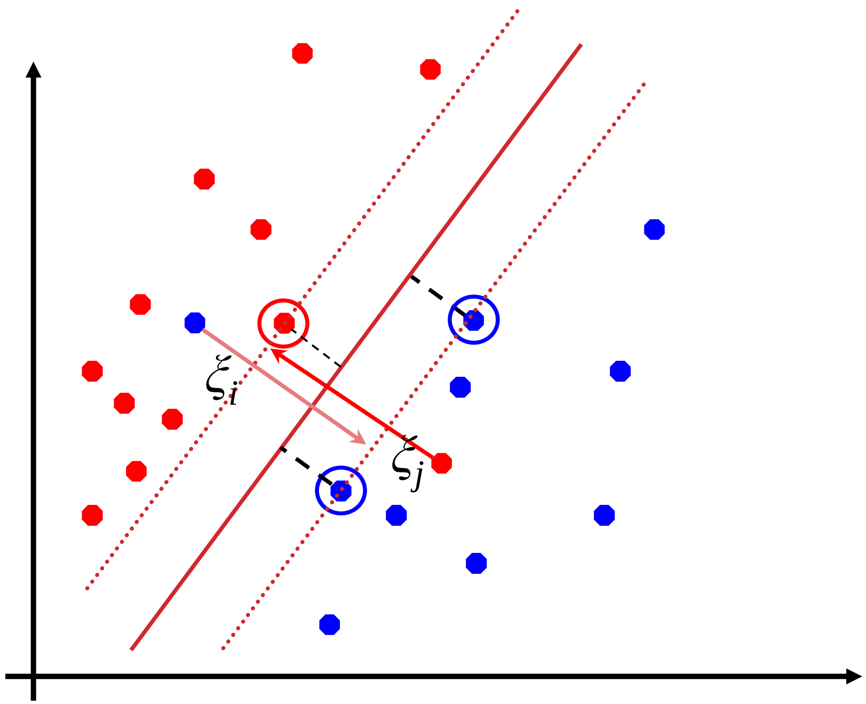
THE KERNEL TRICK



$$\begin{array}{rccl} \phi: & \Re^2 & \longrightarrow & \Re^3 \\ & (x_1, x_2) & \longmapsto & (z_1, z_2, z_3) = (x_1^2, \sqrt{2}x_1x_2, x_2^2) \end{array}$$

[http://www.cs.berkeley.edu/~jordan/courses/281B-spring04/lectures/lec3.pdf]

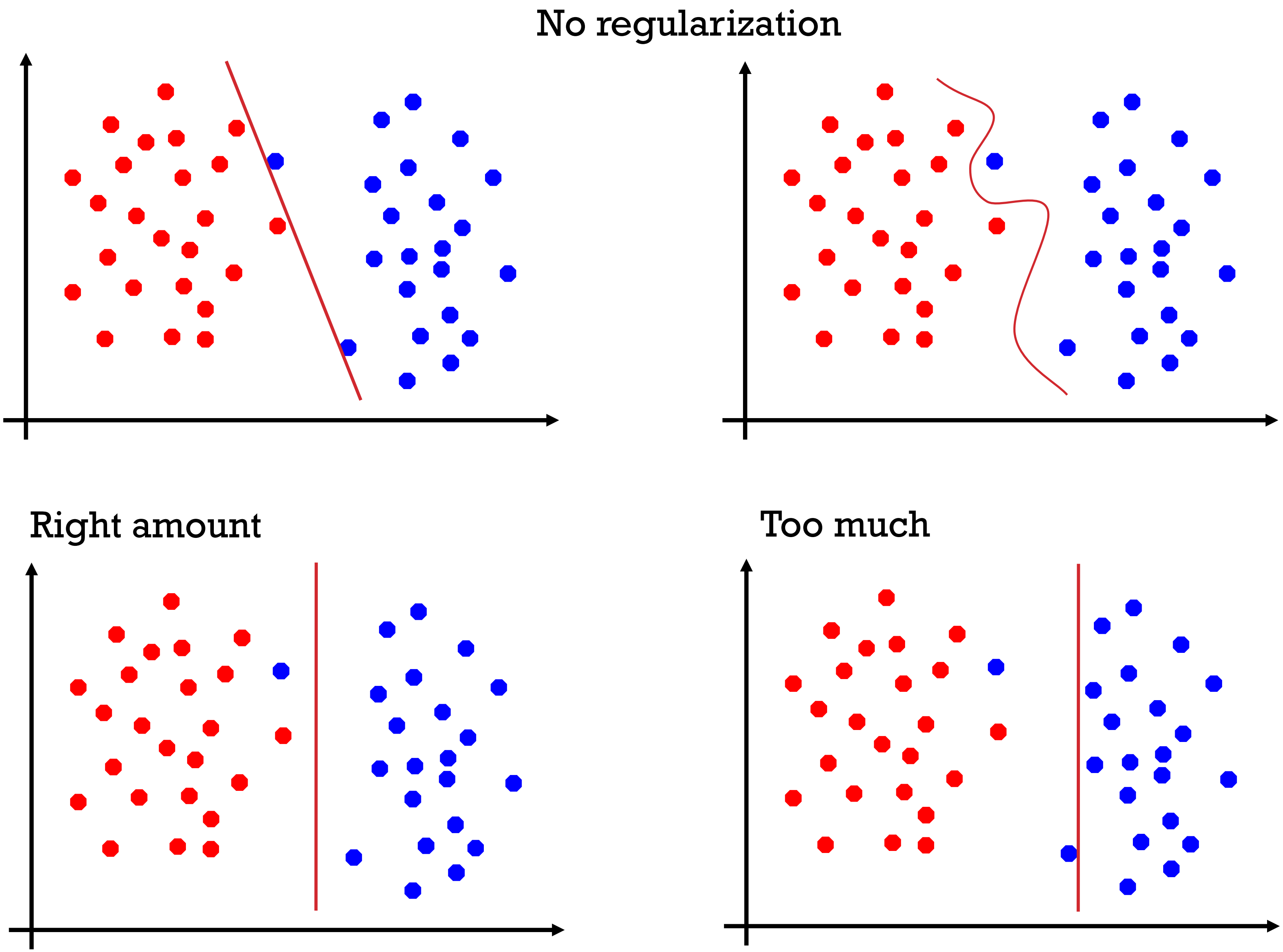
SOFT MARGIN CLASSIFICATION

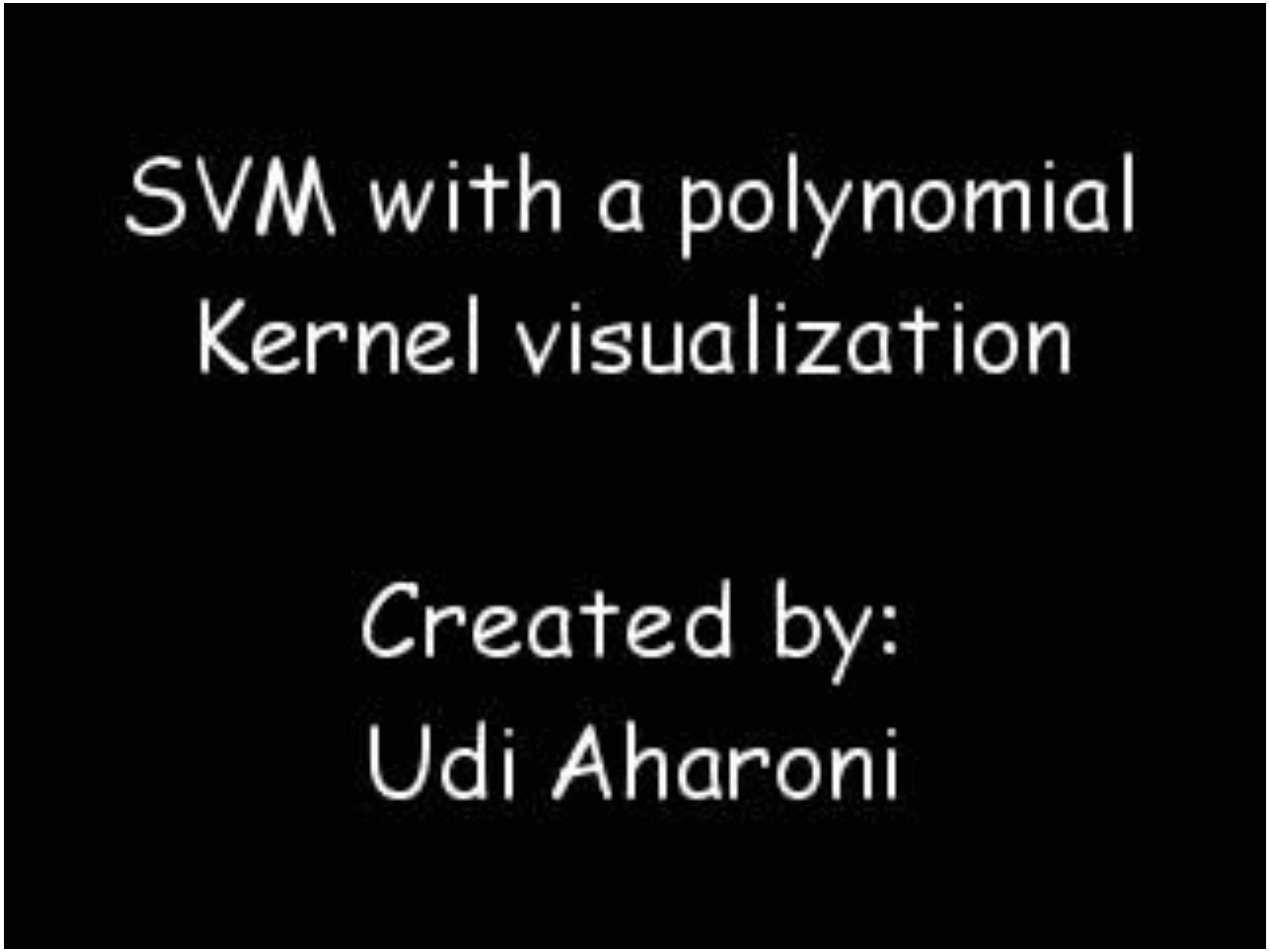


If the training data is not linearly separable, *slack variables* ξ_i (a regularization parameter) can be added to allow misclassification of difficult or noisy examples.

Still, try to minimize training set errors, and to place hyperplane "far" from each class (large margin)

THE IMPACT OF REGULARIZATION





https://www.youtube.com/watch?v=3liCbRZPrZA

MANY CLASSIFIERS TO CHOOSE FROM

K-nearest neighbor

Support Vector Machines

Decision Trees

Random Forrest

(Gradient) Boosted Decision Trees

Logistic Regression

Naïve Bayes

Bayesian network

RBMs

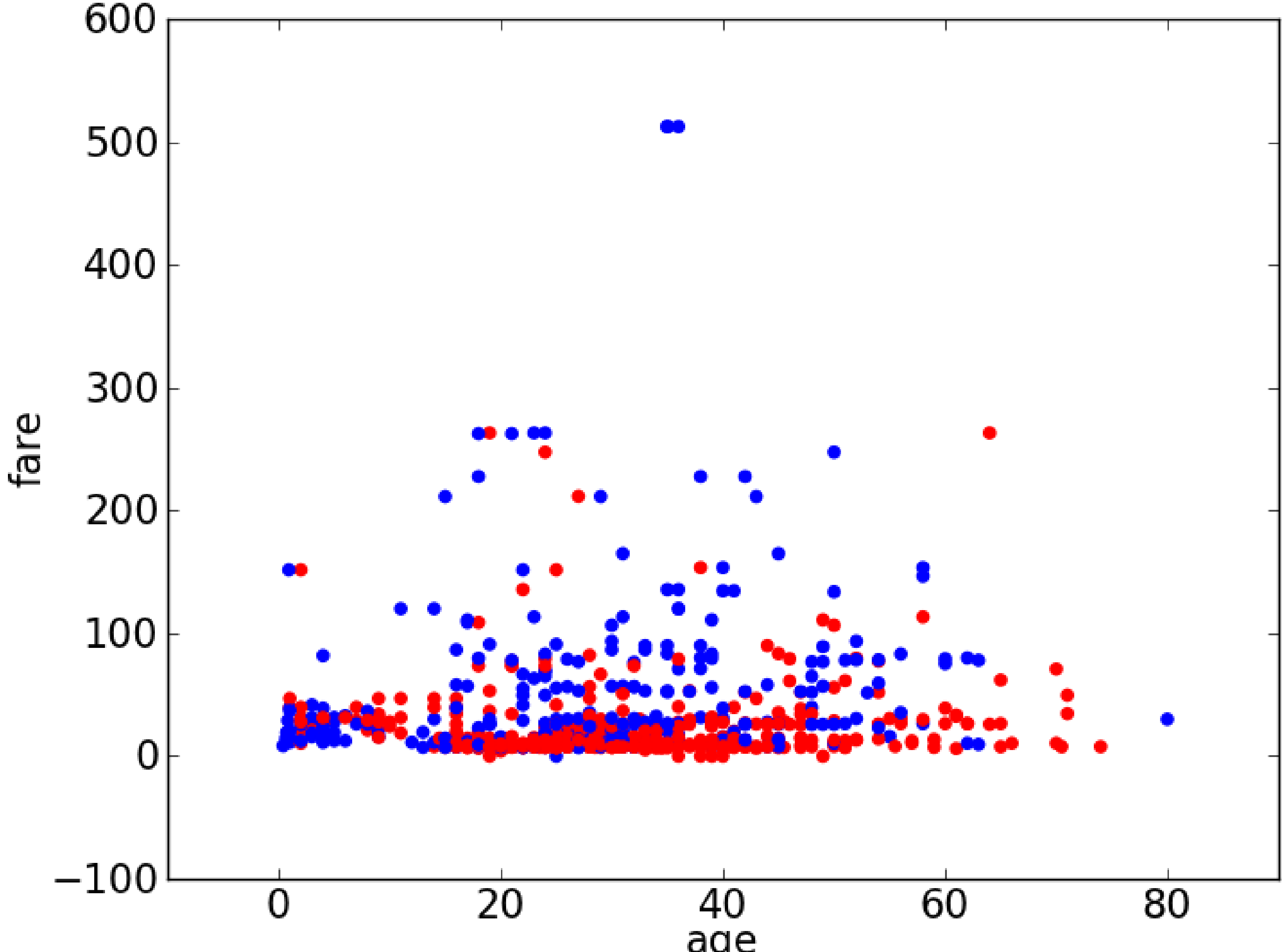
….

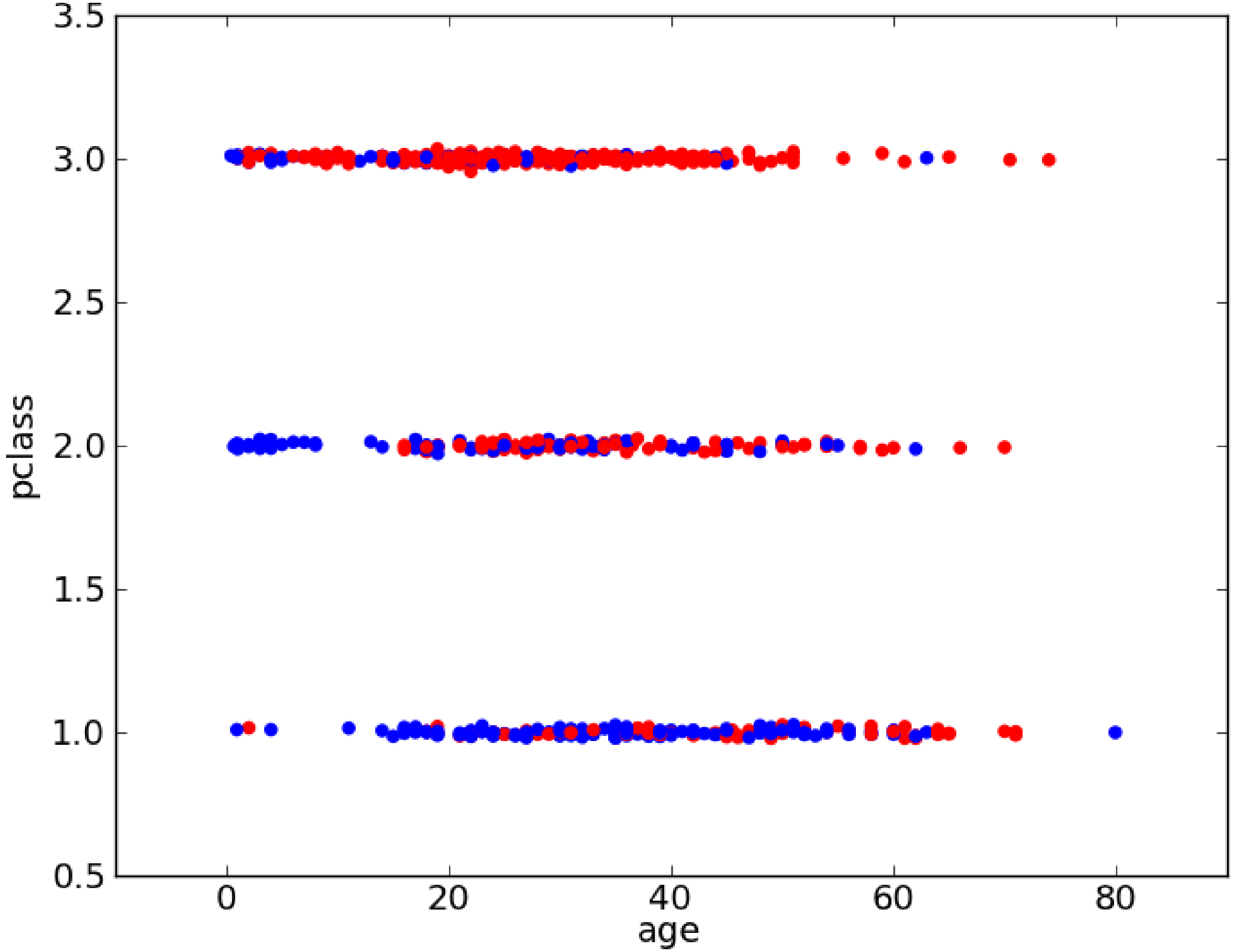
Which is the best one?

Slide credit: D. Hoiem

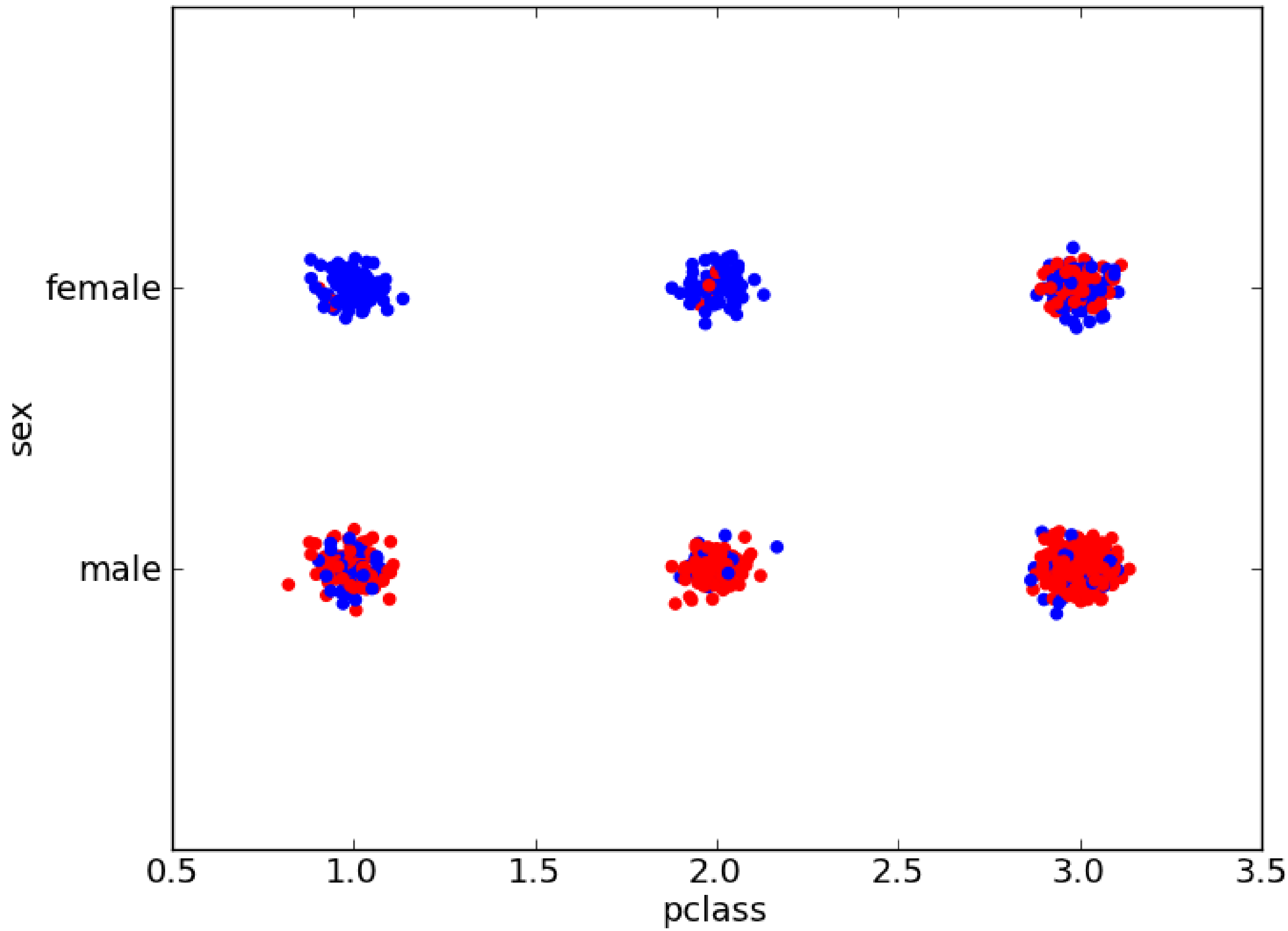
TITANIC DATASET

survived	pclass	sex	age	sibsp	parch	fare	cabin	embarked
0	3	male	22	1	0	7.25		S
1	1	female	38	1	0	71.2833	C85	C
1	3	female	26	0	0	7.925		S
1	1	female	35	1	0	53.1	C123	S
0	3	male	35	0	0	8.05		S
0	3	male		0	0	8.4583		Q
0	1	male	54	0	0	51.8625	E46	S
0	3	male	2	3	1	21.075		S
1	3	female	27	0	2	11.1333		S
1	2	female	14	1	0	30.0708		C
1	3	female	4	1	1	16.7	G6	S
1	1	female	58	0	0	26.55	C103	S
0	3	male	20	0	0	8.05		S


600
500
400
300
200
100
0
−100
fare
0
20
40
60
80
age



3.5
3.0
2.5
2.0
1.5
1.0
0.5
pclass
0
20
40
60
80
age


female
male
sex
0.5
1.0
1.5
2.0
2.5
3.0
3.5
pclass

IF sex='female' THEN survive=yes

ELSE IF sex='male' THEN survive = no

```
confusion matrix

no   yes    <-- classified as
468  109  |   no
 81  233  |  yes
```

(468 + 233) / (468+109+81+233) = 79% correct (and 21% incorrect)

Not bad!

IF pclass='1' THEN survive=yes

ELSE IF pclass='2' THEN survive=yes

ELSE IF pclass='3' THEN survive=no

```
confusion matrix

no   yes    <-- classified as
372  119  |   no
177  223  |  yes
```

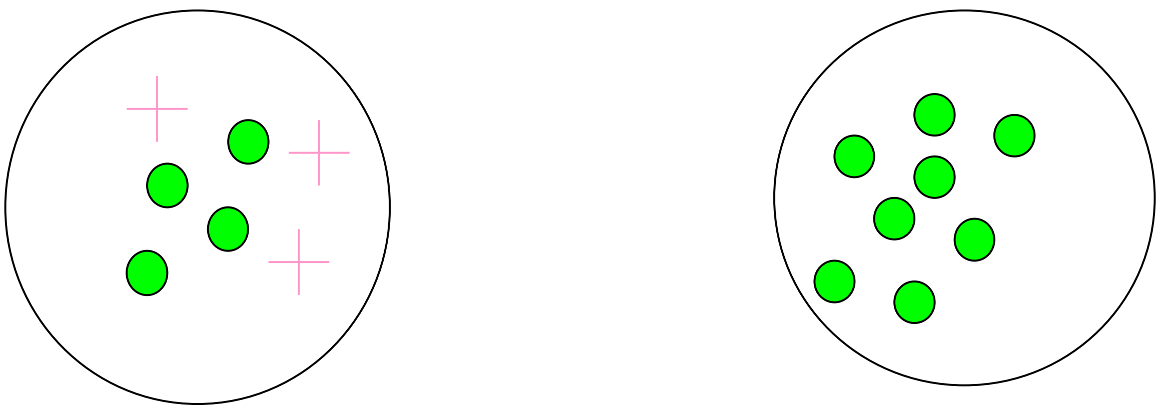
(372 + 223) / (372+119+223+177) = 67% correct (and 33% incorrect)

a little worse

ASIDE ON ENTROPY

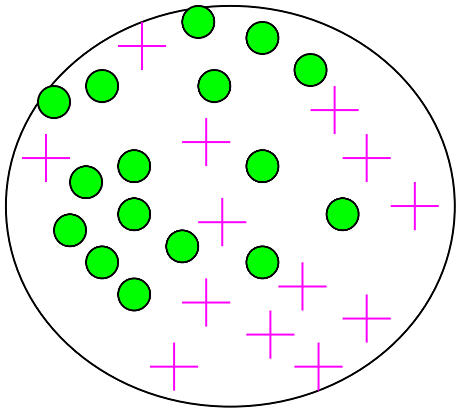
Impurity/Entropy (informal)

– Measures the level of **impurity** in a group of examples

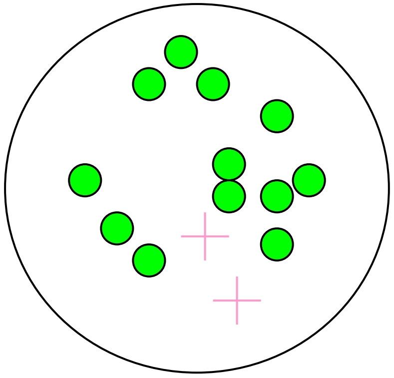


IMPURITY

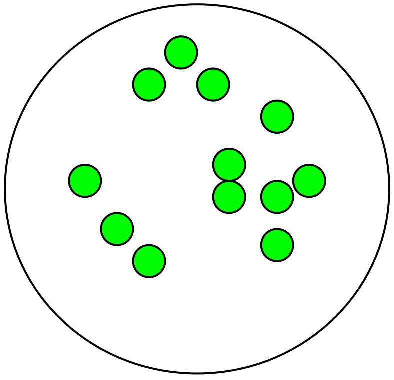
Very impure group



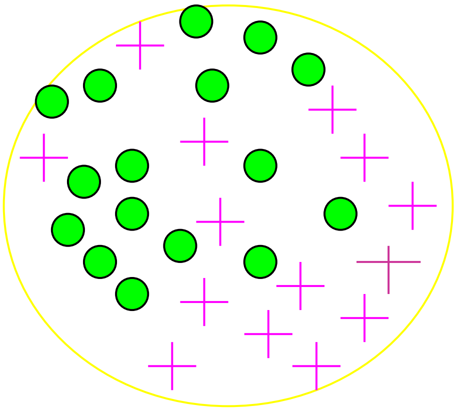
Less impure



Minimum impurity



ENTROPY



- Entropy = $\sum_i - p_i \log_2 p_i$

  p_i is the probability of class i

  Compute it as the proportion of class i in the set.

  16/30 are green circles; 14/30 are pink crosses
  $\log_2(16/30)$ = -.9; $\log_2(14/30)$ = -1.1
  Entropy = -(16/30)(-.9) –(14/30)(-1.1) = .99

- Entropy comes from information theory. The higher the entropy the more the information content.

  What does that mean for learning from examples?

CLICKER

What is the entropy if all examples belong to the same class?

a) 0

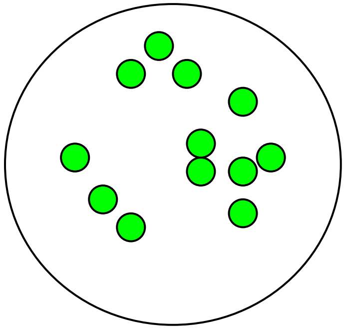
b) 1

c) Infinite

2 CLASS EXAMPLE

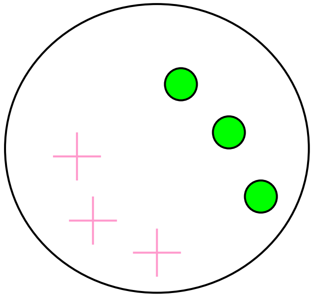
- What is the entropy of a group in which all examples belong to the same class?

Minimum impurity



- What is the entropy of a group with 50% in either class?

Maximum impurity



EXAMPLE: ROLLING A DIE

$$p_1 = \frac{1}{6},\ p_2 = \frac{1}{6},\ p_3 = \frac{1}{6},\ \ldots$$

$$\begin{aligned}\text{Entropy} &= -\sum_i p_i \log_2 p_i \\ &= -6 \times \left(\frac{1}{6}\log_2 \frac{1}{6}\right) \\ &\approx 2.58\end{aligned}$$

CLICKER

Has an unfair/weighted die a higher or lower entropy?

A) Higher

B) Lower

EXAMPLE: ROLLING A WEIGHTED DIE

$$p_1 = 0.1,\ p_2 = 0.1,\ p_3 = 0.1,\ \ldots\ p_6 = 0.5$$

$$\text{Entropy} = -\sum_i p_x \log_2 p_x$$

$$= -5 \times (0.1 \log_2 0.1) - 0.5 \log_2 0.5$$

$$= 2.16$$

The weighted die is has less uncertainty than a fair die

HOW UNCERTAIN IS YOUR DATA?

342/891 survivors in titanic training set

$$-\left(\frac{342}{891}\log_2\frac{342}{891}+\frac{549}{891}\log_2\frac{549}{891}\right)=0.96$$

Say there were only 50 survivors

$$-\left(\frac{50}{891}\log_2\frac{50}{891}+\frac{841}{891}\log_2\frac{841}{891}\right)=0.31$$

IN CLASS TASK

How can you use Entropy to build a decision tree.

Discuss with your neighbor(s)

Discuss the following ideas

Select the feature based on the highest entropy

Select the feature based on the lowest entropy

Stop splitting if the entropy is 0

Select the feature based on the entropy after the split

What if one group is under-/over represented

BACK TO DECISION TREES

Which attribute do we choose at each level?

The one with the highest information gain

- The one that reduces the uncertainty/impurity the most

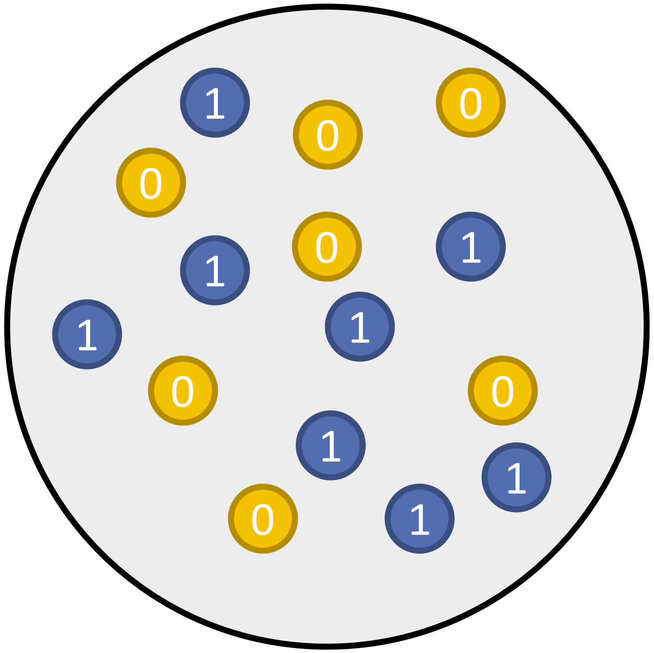
We want to determine which attribute in a given set of training feature vectors is most useful for discriminating between the classes to be learned.

Information gain tells us how important a given attribute of the feature vectors is.

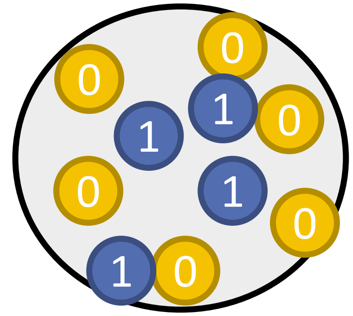
We will use it to decide the ordering of attributes in the nodes of a decision tree.

INFORMATION GAIN

Titanic Entropy = 0.96

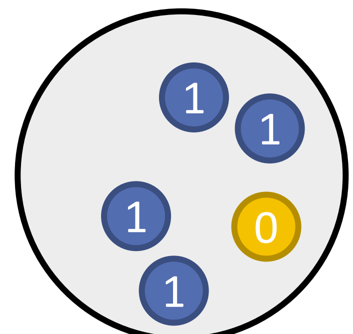


male



Entropy= - 682/843 log(682/843)
- 161/843 log(161/843)
= 0.21

female



Entropy= - 127/466 log(127/466)
- 339/466 log(339/466)
= 0.25

Weighted Entropy: 466/1309 * 0.25 + 843 / 1309 * 0.21 = 0.22
Information Gain for split: 0.96 – 0.22 = 0.74

outlook	temperature	humidity	windy	play
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Before: 14 records, 9 are “yes”

$$-\left(\frac{9}{14}\log_2\frac{9}{14}+\frac{5}{14}\log_2\frac{5}{14}\right)=0.94$$

If we choose **outlook**:

overcast : 4 records, 4 are “yes”

$$-\left(\frac{4}{4}\log_2\frac{4}{4}\right)=0$$

rainy : 5 records, 3 are “yes”

$$-\left(\frac{3}{5}\log_2\frac{3}{5}+\frac{2}{5}\log_2\frac{2}{5}\right)=0.97$$

sunny : 5 records, 2 are “yes”

$$-\left(\frac{2}{5}\log_2\frac{2}{5}+\frac{3}{5}\log_2\frac{3}{5}\right)=0.97$$

Expected new entropy:

$$\frac{4}{14}\times 0.0+\frac{5}{14}\times 0.97+\frac{5}{14}\times 0.97$$

= <u>0.69</u>

outlook	temperature	humidity	windy	play
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Before: 14 records, 9 are "yes"

$$-\left(\frac{9}{14}\log_2\frac{9}{14}+\frac{5}{14}\log_2\frac{5}{14}\right)=0.94$$

Clicker:

If we choose windy, what is the expected entropy?

a) 0.81

= - (6/8 log(6/8) + 2/8 log(2/8))

b) 0.89

= 6/14 * 1 +

(-8/14* (6/8 log(6/8) + 2/8 log(2/8)))

c) 1

= - (0.5 * log(0.5)+0.5 * log(0.5))

outlook	**temperature**	**humidity**	**windy**	**play**
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Before: 14 records, 9 are “yes”

$$-\left(\frac{9}{14}\log_2\frac{9}{14}+\frac{5}{14}\log_2\frac{5}{14}\right)=0.94$$

If we choose **windy**:

FALSE: 8 records, 6 are “yes”

0.81 = -(6/8*log(6/8)+2/8*log(2/8))

TRUE: 6 records, 3 are “yes”

1

Expected new entropy:

0.81(8/14) + 1 (6/14)

= <u>0.89</u>

outlook	temperature	humidity	windy	play
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Before: 14 records, 9 are “yes”

$$-\left(\frac{9}{14}\log_2\frac{9}{14}+\frac{5}{14}\log_2\frac{5}{14}\right)=0.94$$

If we choose **temperature**:

cool : 4 records, 3 are “yes”

0.81

rainy : 4 records, 2 are “yes”

1.0

sunny : 6 records, 4 are “yes”

0.92

Expected new entropy:

0.81(4/14) + 1.0(4/14) + 0.92(6/14)

= <u>0.91</u>

outlook	temperature	humidity	windy	play
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Before: 14 records, 9 are “yes”

$$-\left(\frac{9}{14}\log_2\frac{9}{14}+\frac{5}{14}\log_2\frac{5}{14}\right)=0.94$$

If we choose **humidity**:

normal : 7 records, 6 are “yes”

0.59

high : 7 records, 2 are “yes”

0.86

Expected new entropy:

0.59(7/14) + 0.86(7/14)

= 0.725

outlook	temperature	humidity	windy	play
overcast	cool	normal	TRUE	yes
overcast	hot	high	FALSE	yes
overcast	hot	normal	FALSE	yes
overcast	mild	high	TRUE	yes
rainy	cool	normal	TRUE	no
rainy	mild	high	TRUE	no
rainy	cool	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	mild	normal	FALSE	yes
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Before: 14 records, 9 are "yes"

$$-\left(\frac{9}{14}\log_2\frac{9}{14}+\frac{5}{14}\log_2\frac{5}{14}\right)=0.94$$

outlook

0.94 – 0.69 = 0.25 highest gain

temperature

0.94 – 0.91 = 0.03

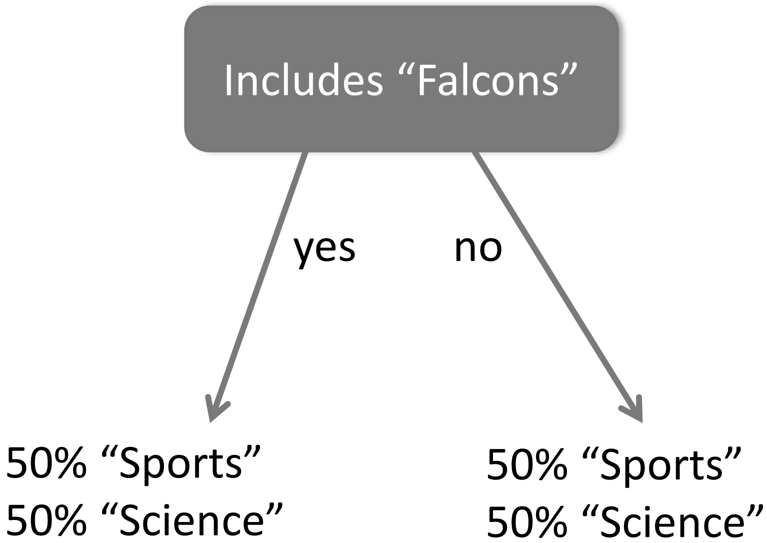
humidity

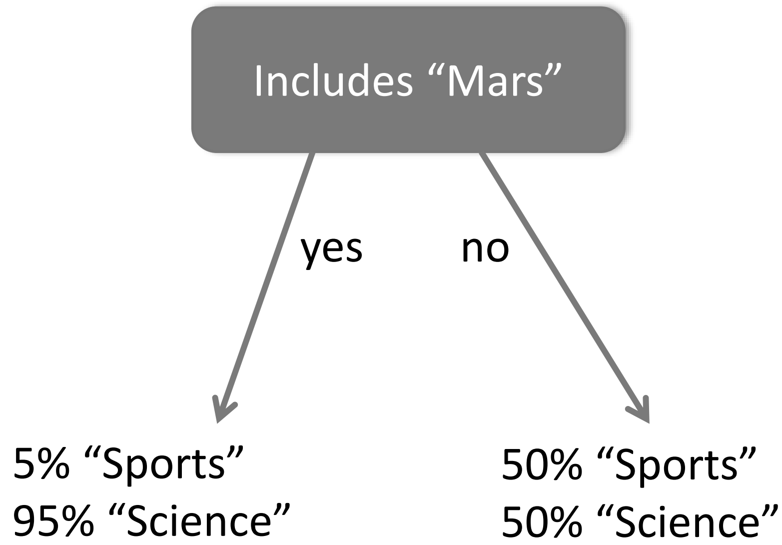
0.94 – 0.725 = 0.215

windy

0.94 – 0.87 = 0.07

DOCUMENT CLASSIFICATION





Clicker Question (assuming equal size):
a) Falcon’s Information Gain is higher
b) Mars’ Information Gain is higher

BUILDING A DECISION TREE (ID3 ALGORITHM)

Assume attributes are discrete

- Discretize continuous attributes

Choose the attribute with the highest Information Gain

Create branches for each value of attribute

Examples partitioned based on selected attributes

Repeat with remaining attributes

Stopping conditions

- All examples assigned the same label
- No examples left

PROBLEMS

Expensive to train

Prone to overfitting

- Drive to perfection on training data, bad on test data
- Pruning can help: remove or aggregate subtrees that provide little discriminatory power (C45)

C4.5 EXTENSIONS

Continuous Attributes

outlook	temperature	humidity	windy	play
overcast	cool	60	TRUE	yes
overcast	hot	80	FALSE	yes
overcast	hot	63	FALSE	yes
overcast	mild	81	TRUE	yes
rainy	cool	58	TRUE	no
rainy	mild	90	TRUE	no
rainy	cool	54	FALSE	yes
rainy	mild	92	FALSE	yes
rainy	mild	59	FALSE	yes
sunny	hot	90	FALSE	no
sunny	hot	89	TRUE	no
sunny	mild	90	FALSE	no
sunny	cool	60	FALSE	yes
sunny	mild	62	TRUE	yes

Consider every possible binary partition; choose the partition with the highest gain

outlook	temperature	humidity	windy	play
rainy	mild	54	FALSE	yes
overcast	hot	58	FALSE	yes
overcast	cool	59	TRUE	yes
rainy	cool	60	FALSE	yes
overcast	mild	60	TRUE	yes
overcast	hot	62	FALSE	yes
rainy	mild	63	TRUE	no
sunny	cool	80	FALSE	yes
rainy	mild	81	FALSE	yes
sunny	mild	89	TRUE	yes
sunny	hot	90	FALSE	no
rainy	cool	90	TRUE	no
sunny	hot	90	TRUE	no
sunny	mild	92	FALSE	no

E(6/6) = 0.0

E(3/8) + E(5/8) = 0.95

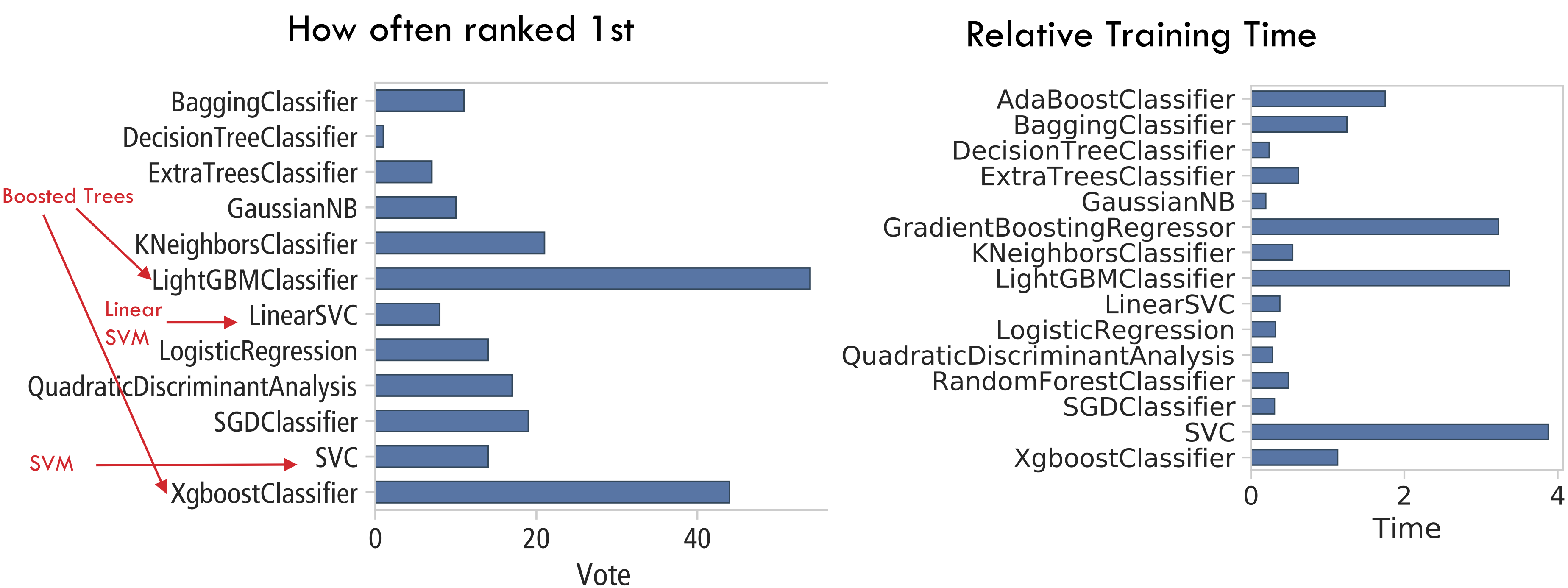
E(9/10)+ E(1/10) = 0.47

E(4/4) = 0.0

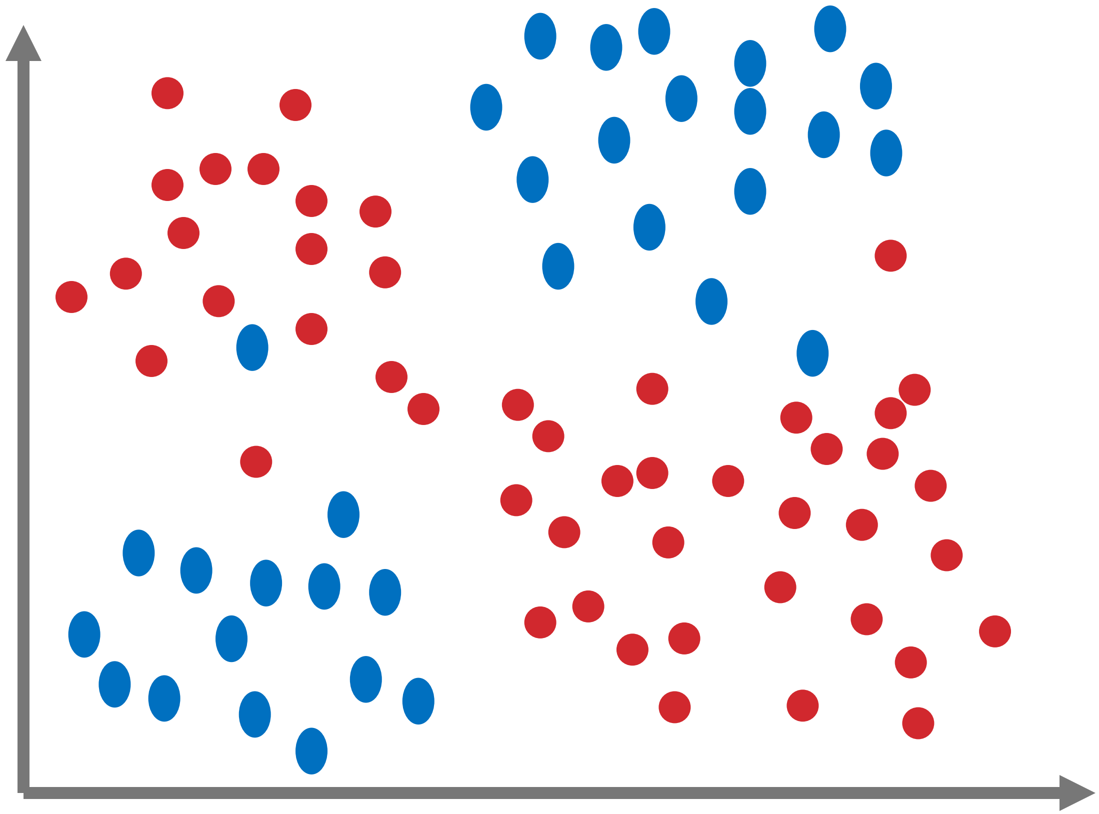
Expect = 8/14*0.95 + 6/14*0 = 0.54

Expect = 10/14*0.47 + 4/14*0 = 0.33

PERFORMANCE OF DIFFERENT ML MODEL FAMILIES



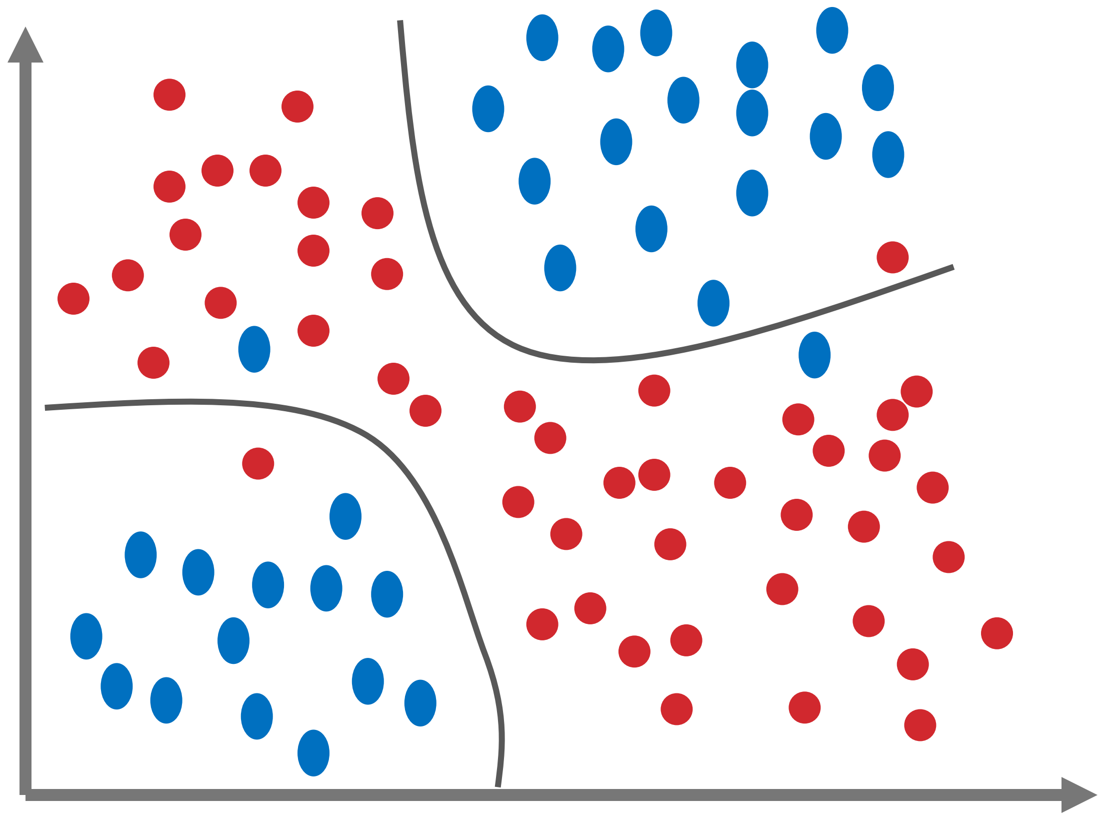
IN-CLASS TASK



How would you draw the expected decision boundary for

- Random Forrest
- SVM w/ kernel and regularization
- 1-KNN

CLICKER



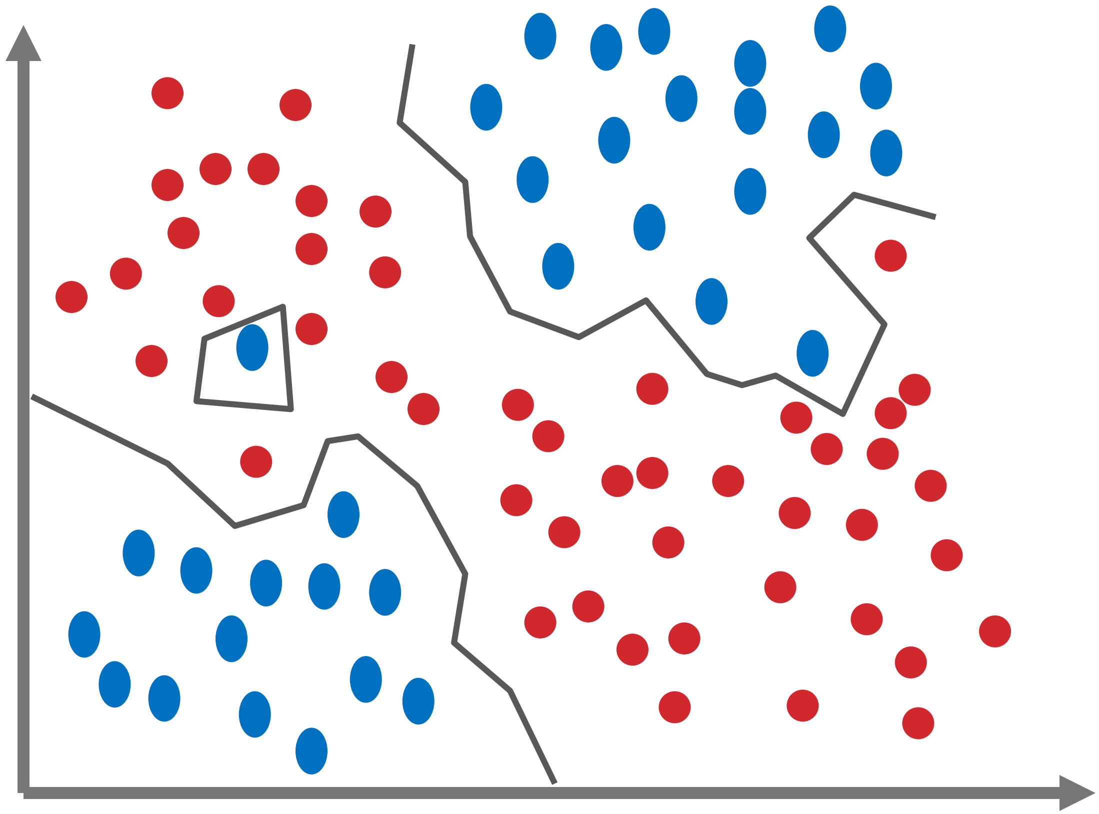
The decision boundary looks like the one of:

a) Random Forrest

b) SVM w/ kernel and regularization

c) 1-KNN

CLICKER



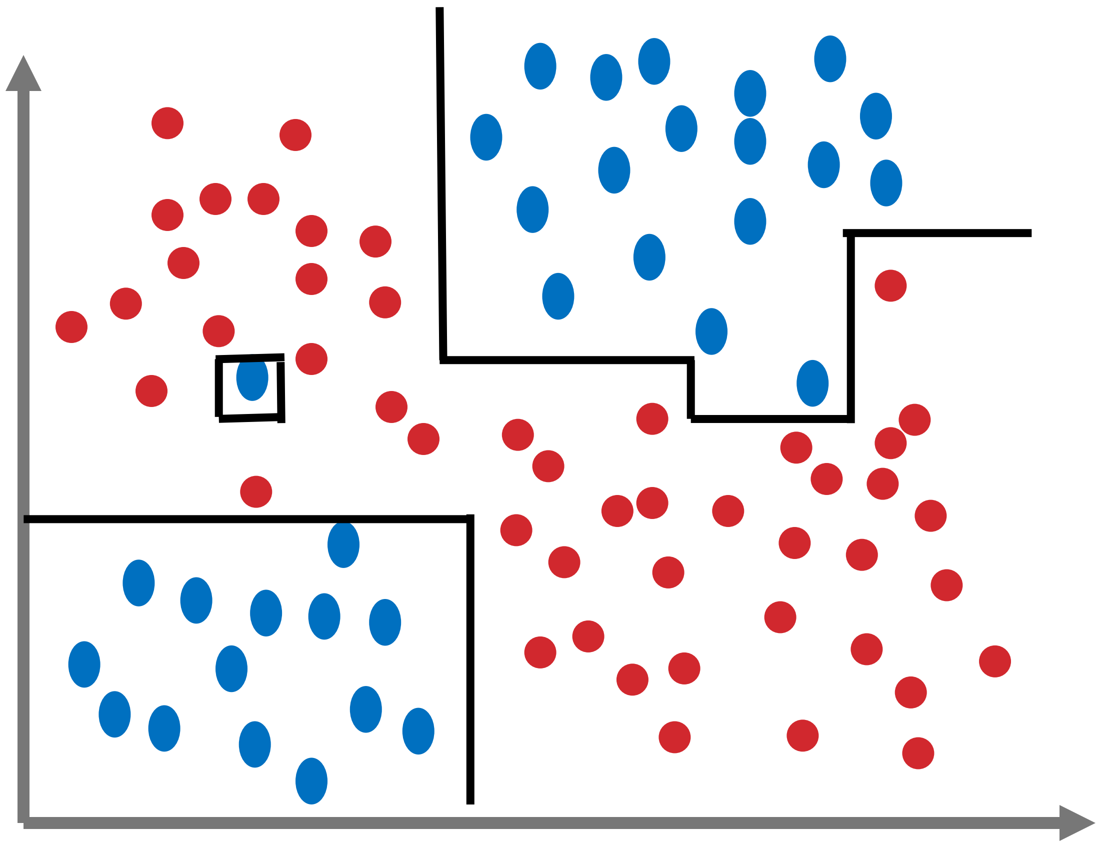
The decision boundary looks like the one of:

a) Random Forrest

b) SVM w/ kernel and regularization

c) 1-KNN

RANDOM FORREST



The decision boundary looks like the one of:

a) Random Forrest

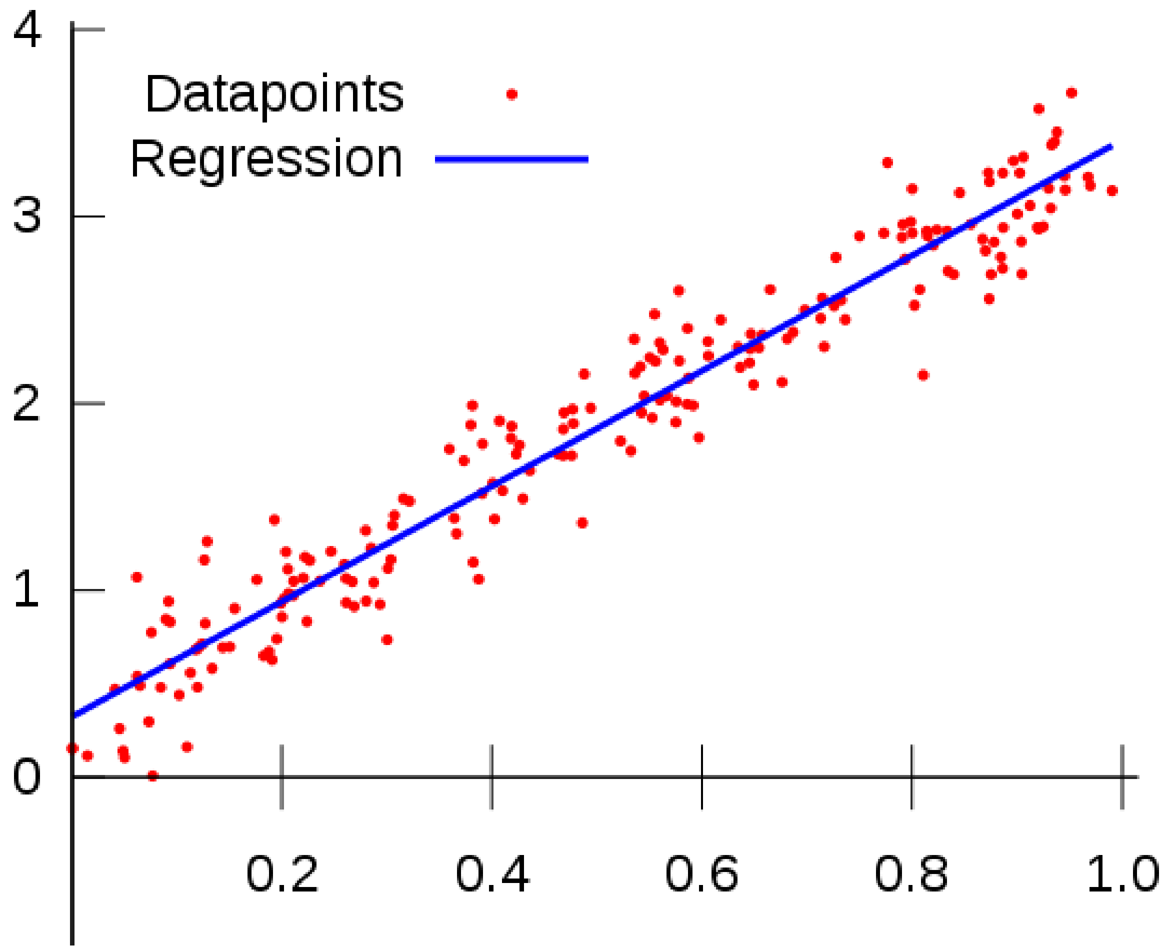
b) SVM w/ kernel and regularization

c) 1-KNN

MACHINE LEARNING PROBLEMS

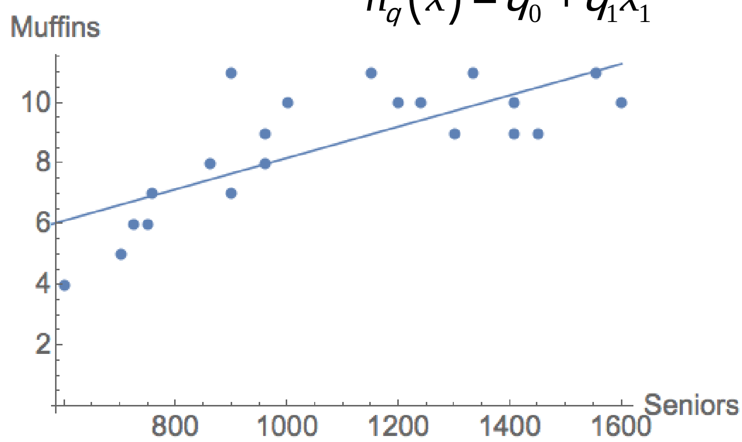
	Supervised Learning	Unsupervised Learning
Discrete	classification or categorization	clustering
Continuous	regression	dimensionality reduction

LINEAR REGRESSION

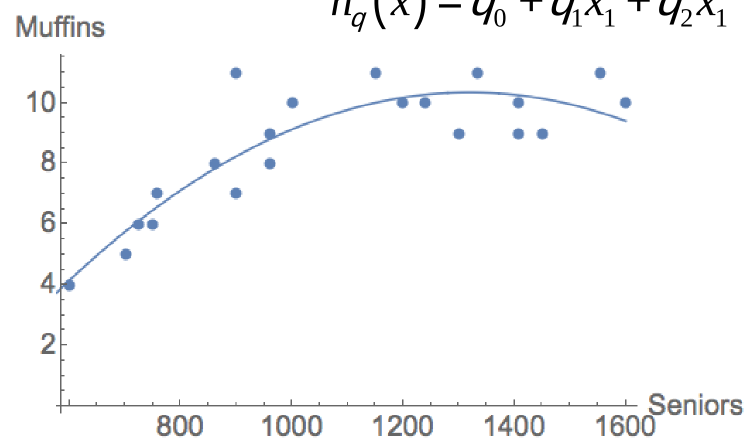


POLYNOMIAL REGRESSION

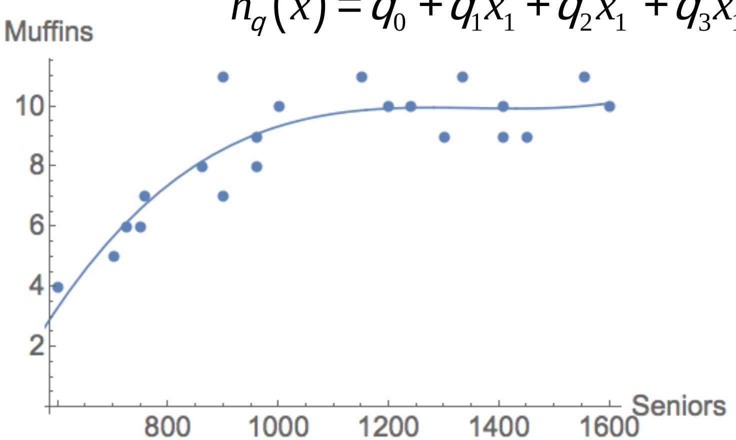
$$h_q(x) = q_0 + q_1 x_1$$



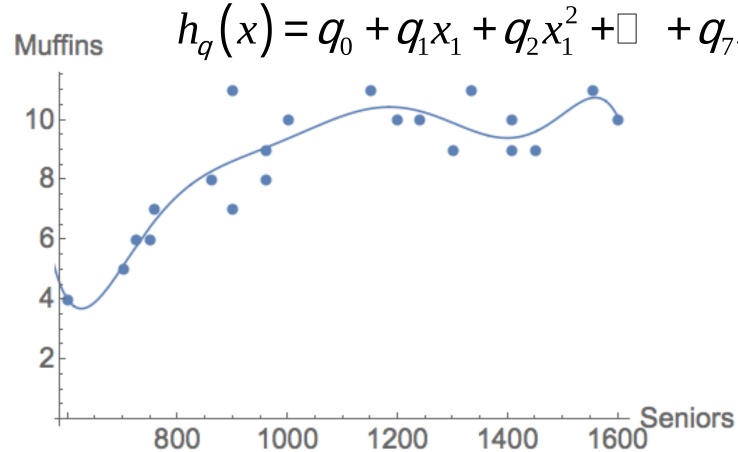
$$h_q(x) = q_0 + q_1 x_1 + q_2 x_1^2$$



$$h_q(x) = q_0 + q_1 x_1 + q_2 x_1^2 + q_3 x_1^3$$



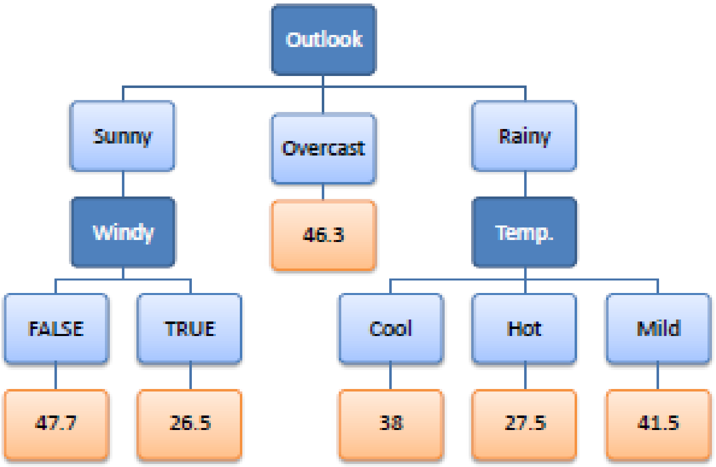
$$h_q(x) = q_0 + q_1 x_1 + q_2 x_1^2 + \cdots + q_7 x_1^7$$



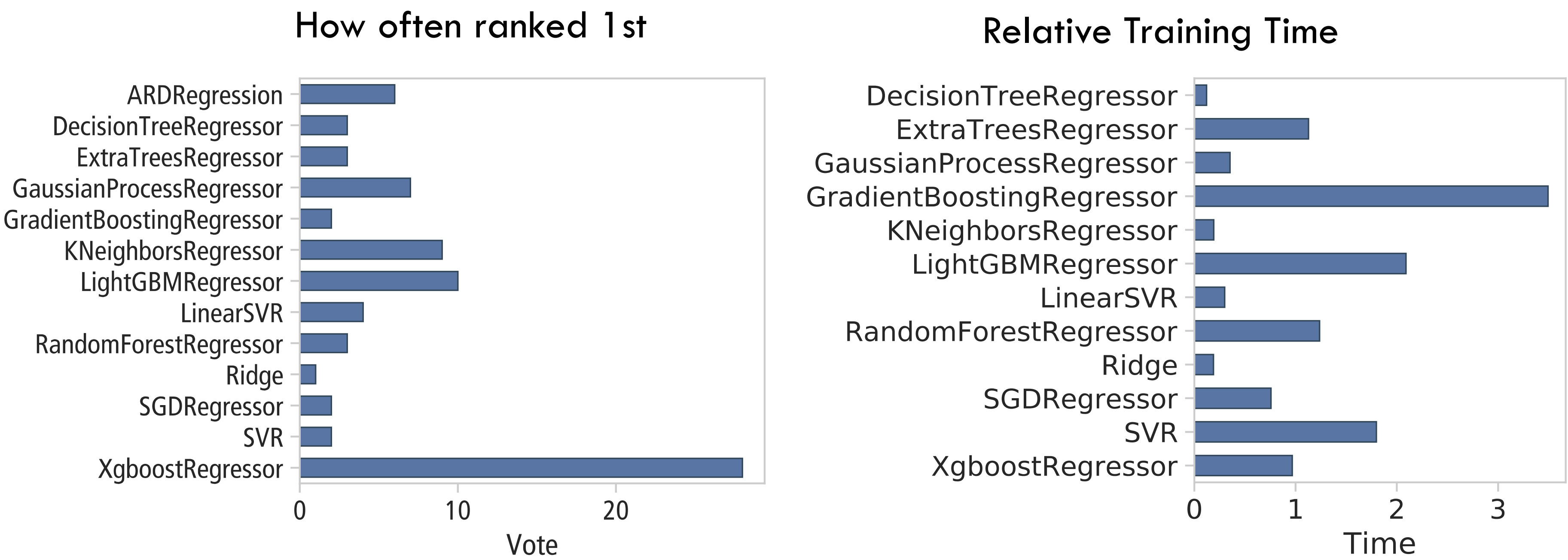
DECISION TREE - REGRESSION

Outlook	Temp.	Humidity	Windy	Hours Played
Rainy	Hot	High	False	26
Rainy	Hot	High	True	30
Overcast	Hot	High	False	48
Sunny	Mild	High	False	46
Sunny	Cool	Normal	False	62
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	36
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	48
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	62
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30

Predictors: Outlook, Temp., Humidity, Windy; Target: Hours Played



PERFORMANCE



Machine Learnin

Nightmare
series

What if you model has an high error?

- Try getting more training examples
- Try smaller sets of features
- Try getting additional features
- Try creating features from existing features (kernels)
- Try decrease regularization
- Try increase regularization

What Error/Quality Metric to use?

Classification:

- **Accuracy**
- **F-score**
- **F1-micro**
- **F1-macro**
- ROC AUC (micro, macro)
- …

Regression

- Mean-Squared Error
- Root-Mean Squared Error
- Mean absolute Error
- R^2
- Cohen Kappa
- ..

Precision, Recall, Accuracy

	True	False
True	tp	fp
False	fn	tn

- **Precision**: correctly identified positive cases

  Precision $P = tp/(tp + fp)$

- **Recall**: correctly identified positive cases from all the actual positive cases.

  Recall $R = tp/(tp + fn)$

- **Accuracy:** measure of all the correctly identified cases

  Accuracy $R = (tp+tn)/(tp + fp + fn + tn)$

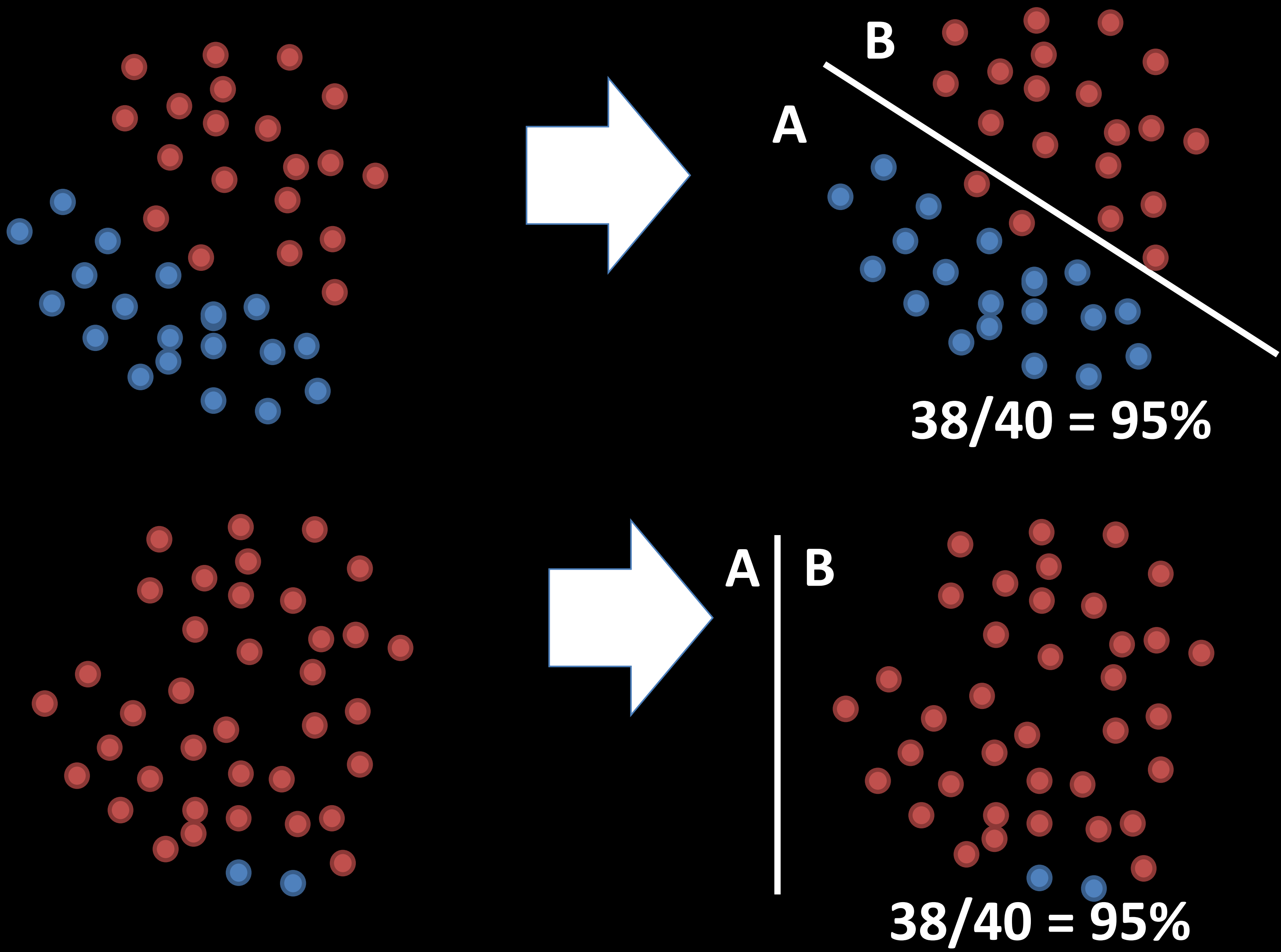
Evaluation: Accuracy isn't always enough

- How do you interpret 90% accuracy?

Evaluation: Accuracy isn’t always enough

- How do you interpret 90% accuracy?
  - You can’t; it depends on the problem
- Need a baseline:
  - Base Rate
    - Accuracy of trivially predicting the most-frequent class
  - Random Rate
    - Accuracy of making a random class assignment
    - Might apply prior knowledge to assign random distribution
  - Naïve Rate
    - Accuracy of some simple default or pre-existing model
    - Ex: “All females survived”

Why Optimize? Pitfalls



What Error/Quality Metric to use?

Classification:

- Accuracy
- **F-score**
- F1-micro
- F1-macro
- ROC AUC (micro, macro)
- …

Regression

- Mean-Squared Error
- Root-Mean Squared Error
- Mean absolute Error
- R^2
- Cohen Kappa
- ..

Precision, Recall, Accuracy

		True Label	
		True	False
Predicted Label	True	tp	fp
	False	fn	tn

- **Precision**: correctly identified positive cases

  Precision P = tp/(tp + fp)

- **Recall**: correctly identified positive cases from all the actual positive cases.

  Recall R = tp/(tp + fn)

- **F-Score:** is the harmonic mean of precision and recall

$$F = \frac{2}{\frac{1}{R}+\frac{1}{P}} = \frac{tp}{tp+\frac{1}{2}(fp+fn)}$$

F1 Micro

		True Label		
		L1	L2	L3
Predicted Label	L1	7	1	4
	L2	0	1	12
	L3	1	6	6

Precision micro: true positives for all the classes divided by the all positive predictions
Precision Score Micro = TP / (TP + FP)
TP = (7 + 1 + 6)
FP = 1 + 4 + 0 + 12 + 1 + 6

Recall micro: Sum of **true positives for all the classes** divided by the actual positives.
Recall Score Micro: TP / (TP + FN)

F1 Score: $\frac{tp}{tp+\frac{1}{2}(fp+fn)}$

Macro

F1 Macro

		True Label		
		L1	L2	L3
Predicted Label	L1	7	1	4
	L2	0	1	12
	L3	1	6	6

Precision micro: **arithmetic mean of all the precision scores** of different classes
Precision Score Macro = ((7/8) + (1/8) + (6/22))/3

Recall micro: **arithmetic mean of all the recall scores** .

When to use F1 Micro and when to use F1 Macro?

F1 Macro

		True Label		
		L1	L2	L3
Predicted Label	L1	7	1	4
	L2	0	1	12
	L3	1	6	6

Precision micro: **arithmetic mean of all the precision scores** of different classes
Precision Score Macro = ((7/8) + (1/8) + (6/22))/3

Recall micro: **arithmetic mean of all the recall scores** .

When to use F1 Micro and when to use F1 Macro?

- Micro weights each instance or prediction equally.
- Macro weights each class equally (better for imbalance of labels)
- Use weighted macro-averaging score in case of class imbalances (different number of instances related to different class labels).

ROC AUC

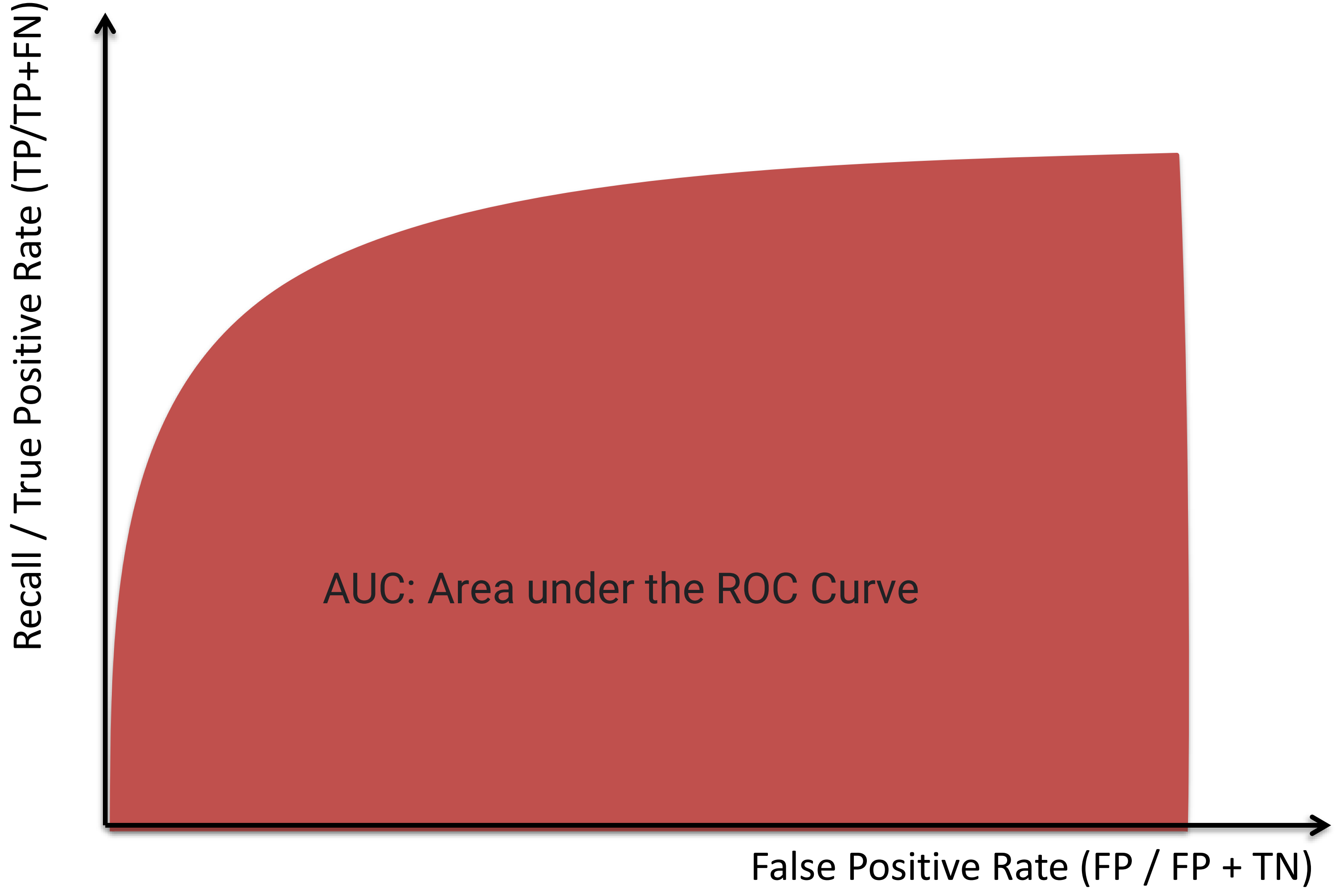
(usually used for models with a threshold)

What would be the ideal ROC curve?

How would a random guess look like

ROC AUC

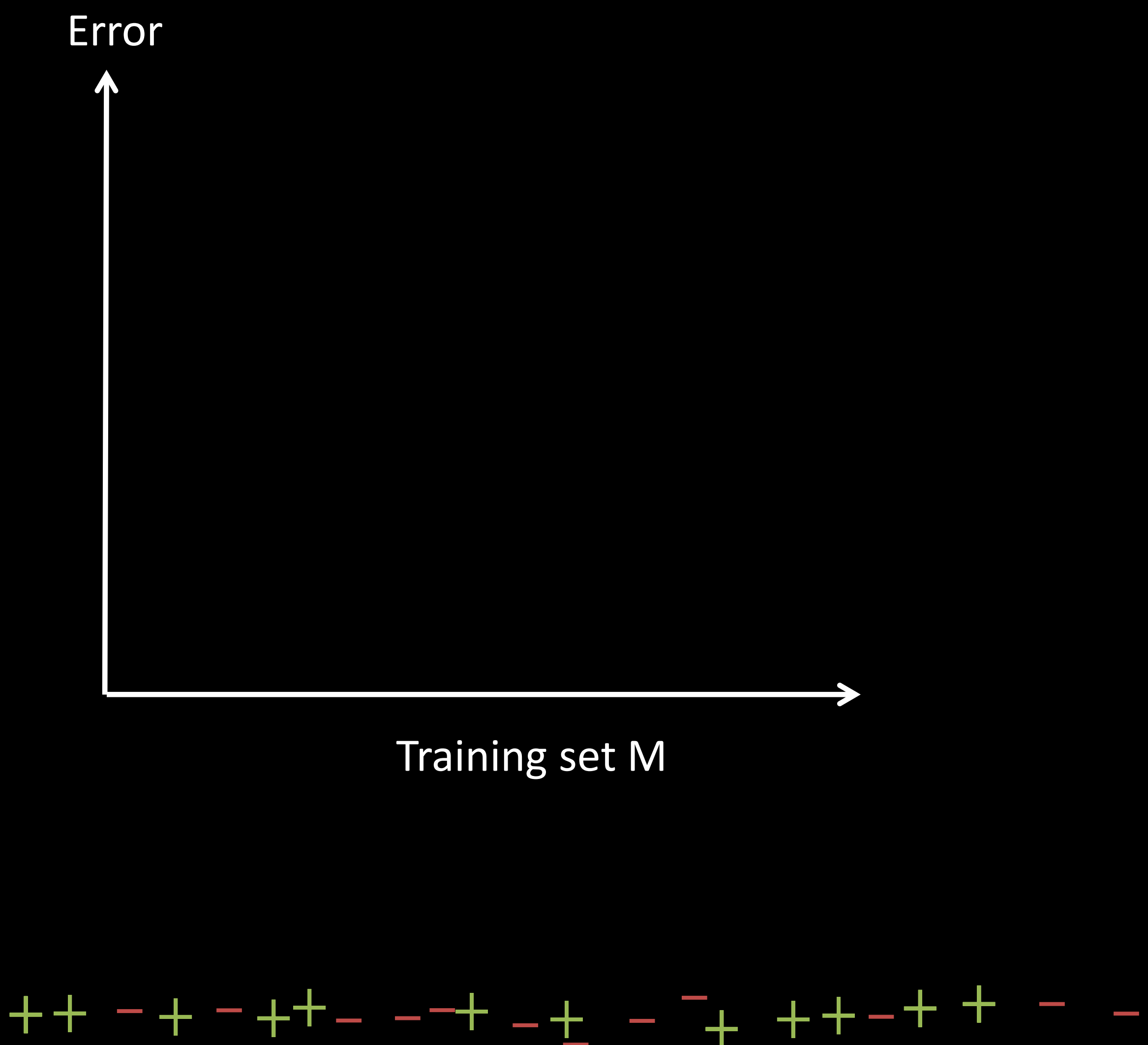
(usually used for models with a threshold)



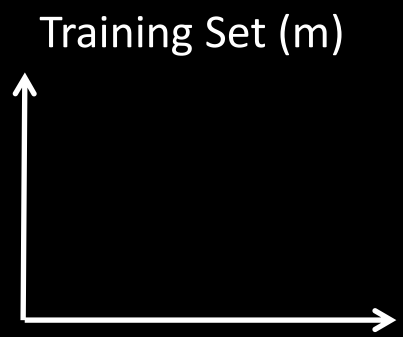
What if your model has a high error?

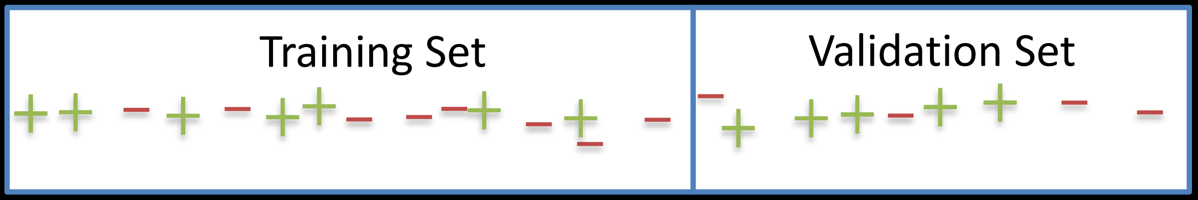
- Try getting more training examples
- Try smaller sets of features
- Try getting additional features
- Try creating features from existing features (kernels)
- Try decrease regularization
- Try increase regularization

Bias and Variance

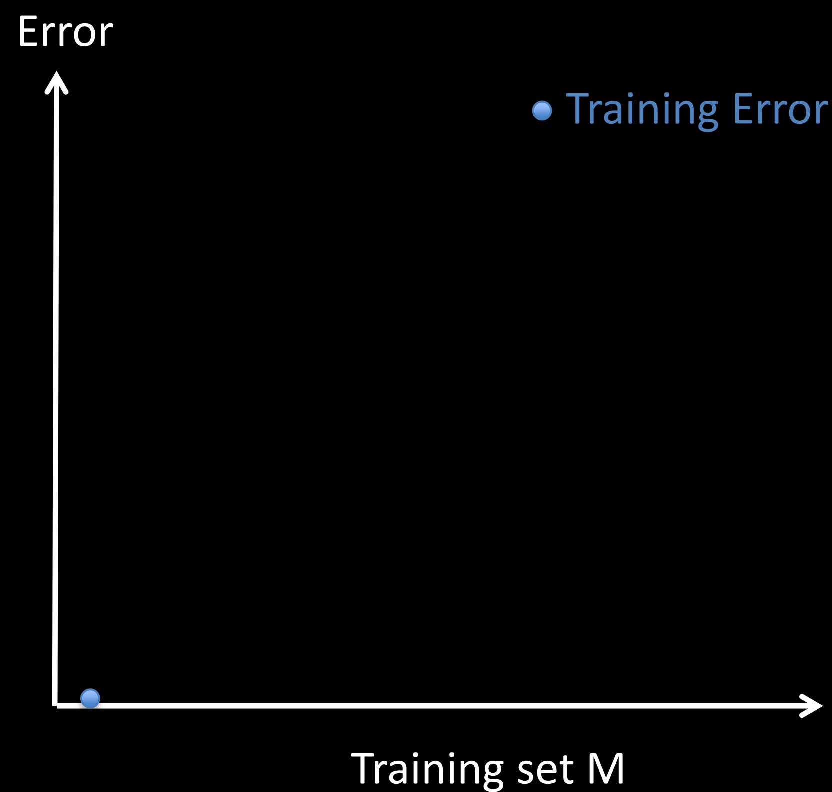


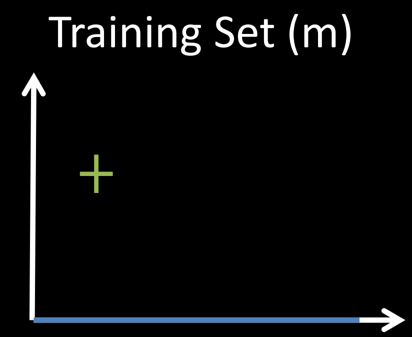
Bias and Variance

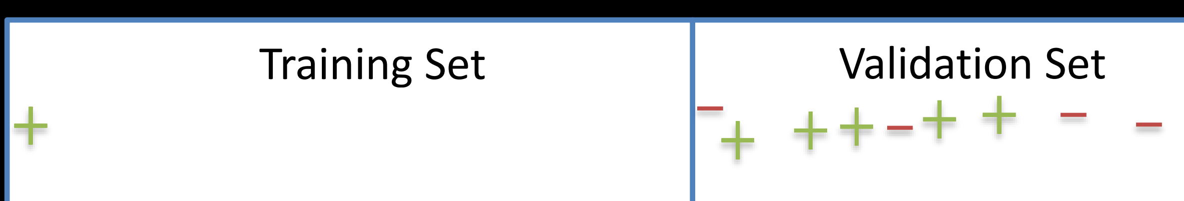




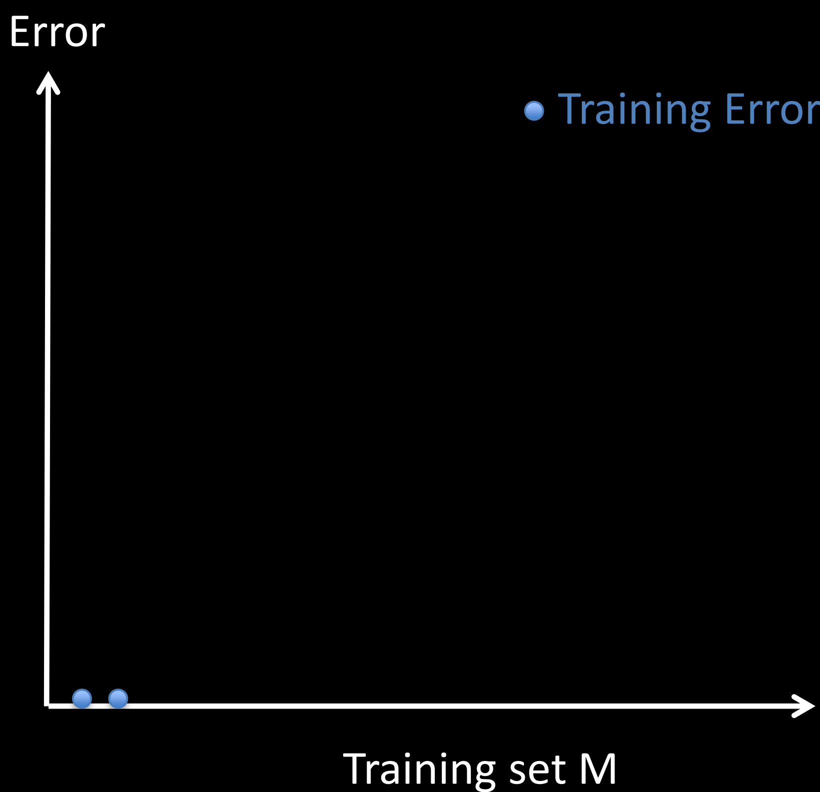
Bias and Variance

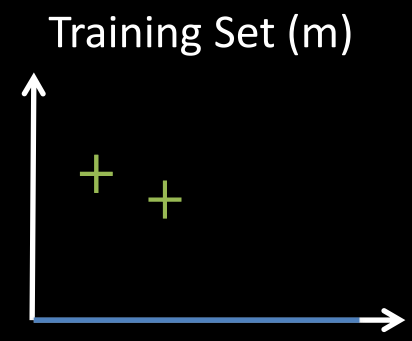




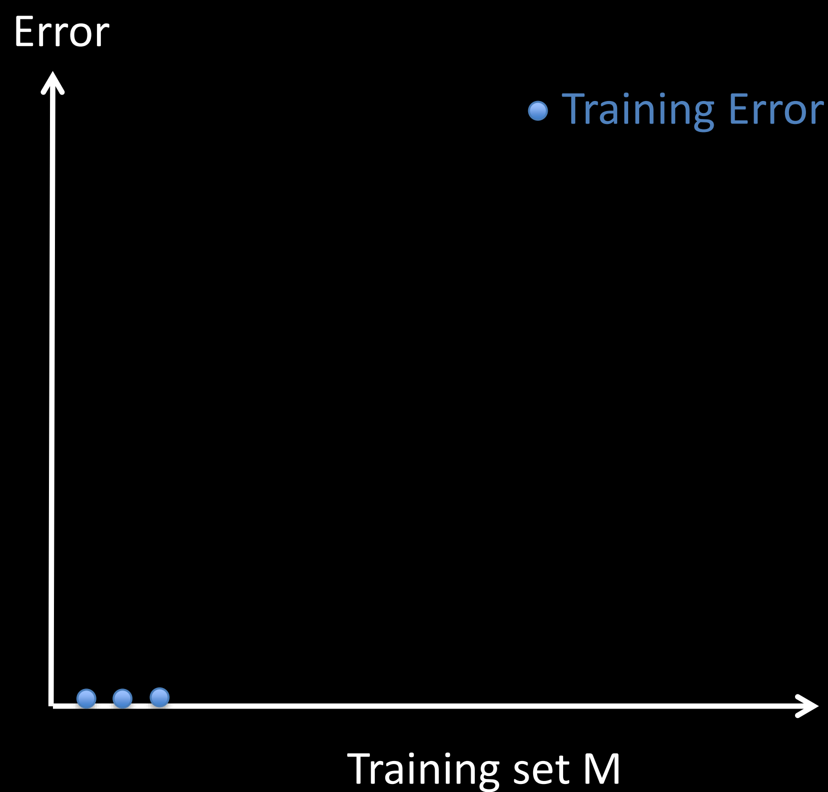


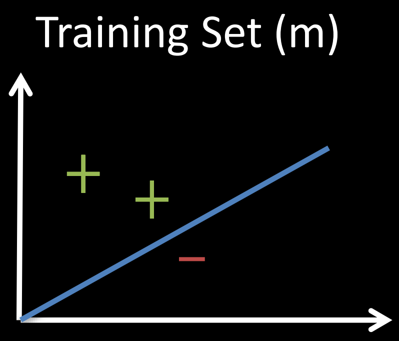
Bias and Variance



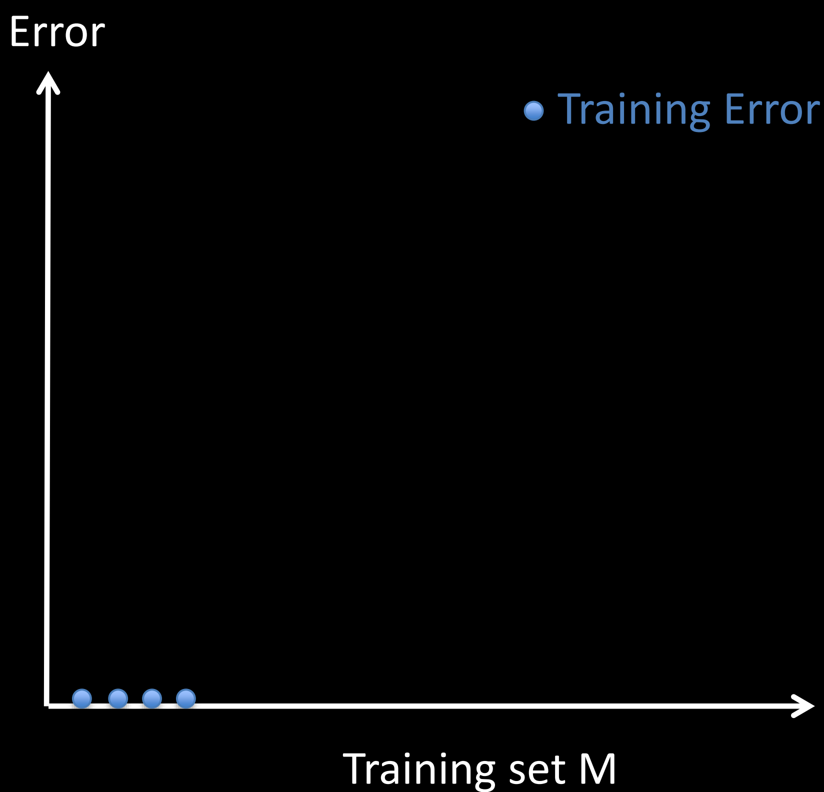


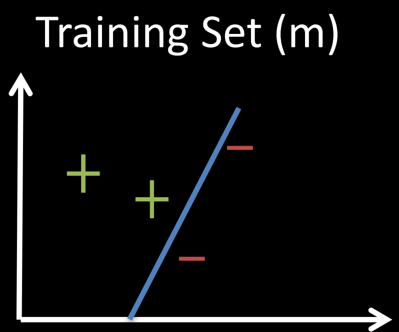
Bias and Variance



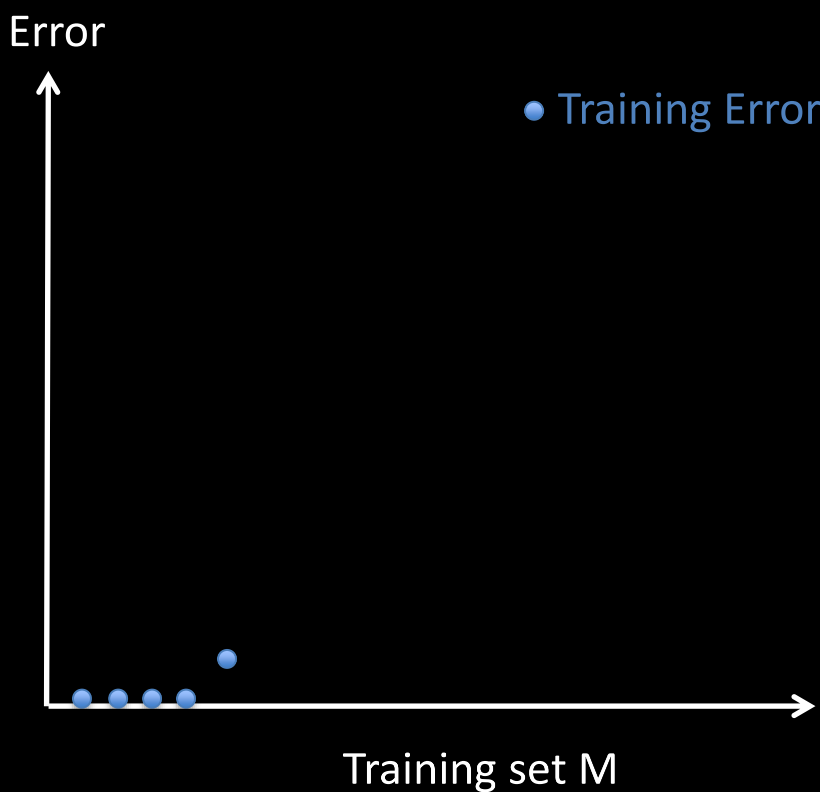


Bias and Variance

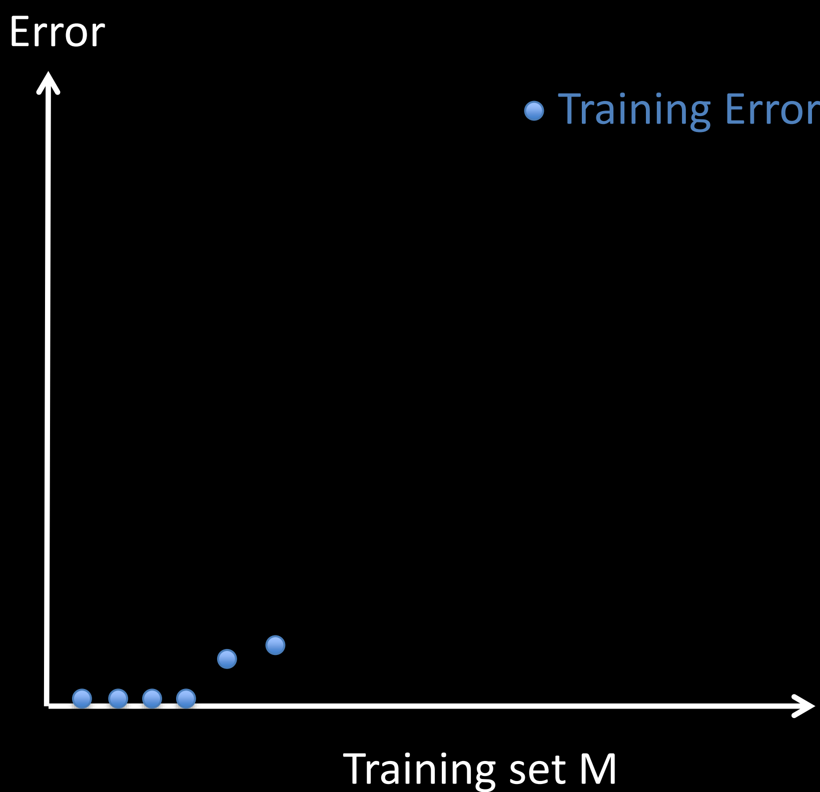


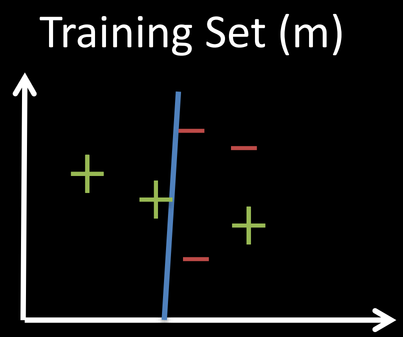


Bias and Variance

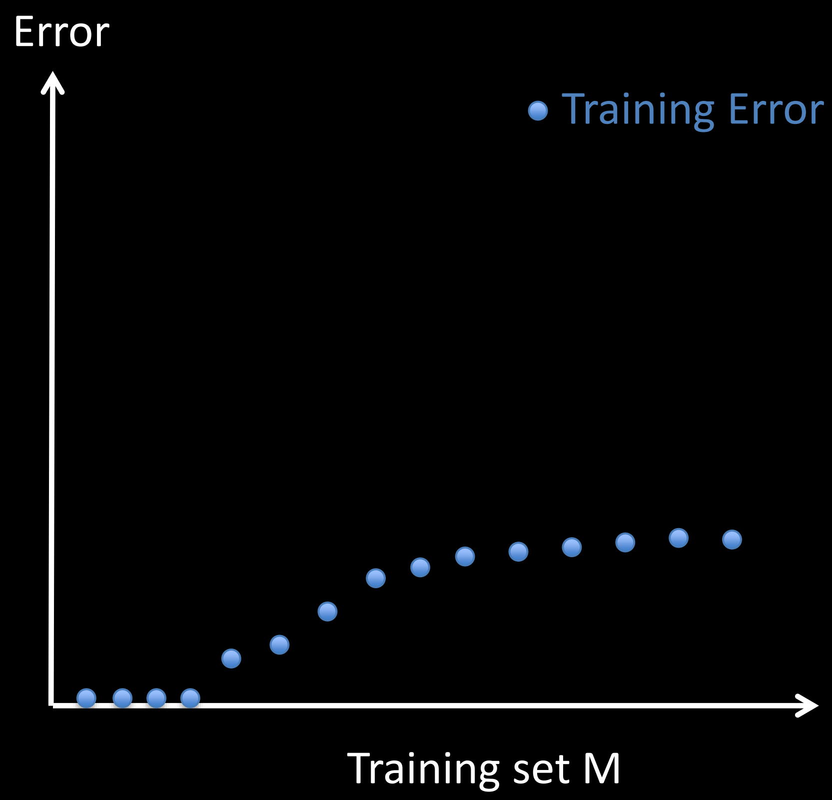


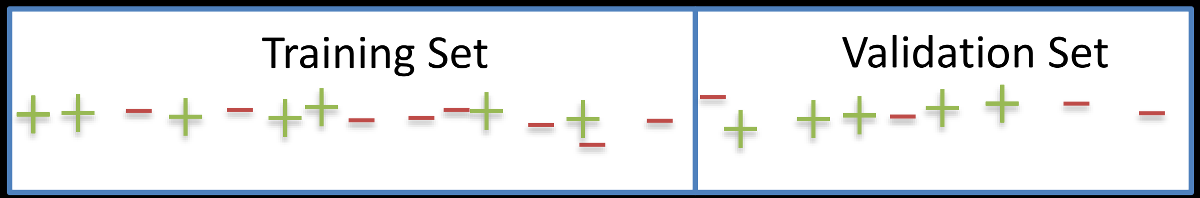
Bias and Variance



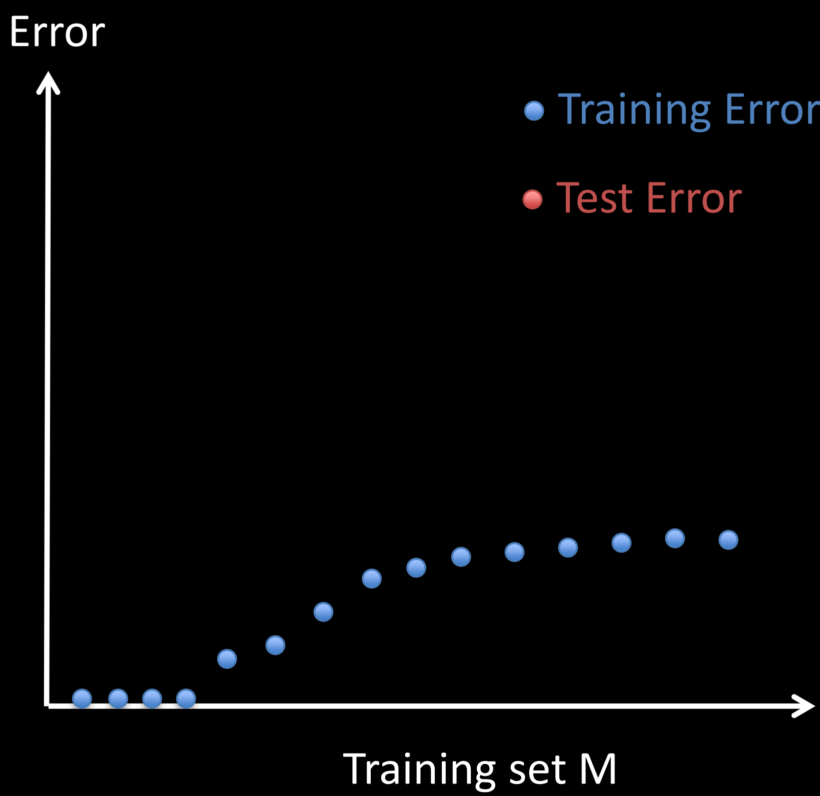


Bias and Variance





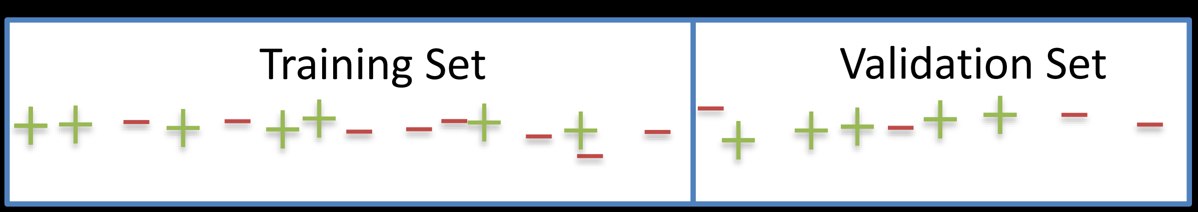
Bias and Variance



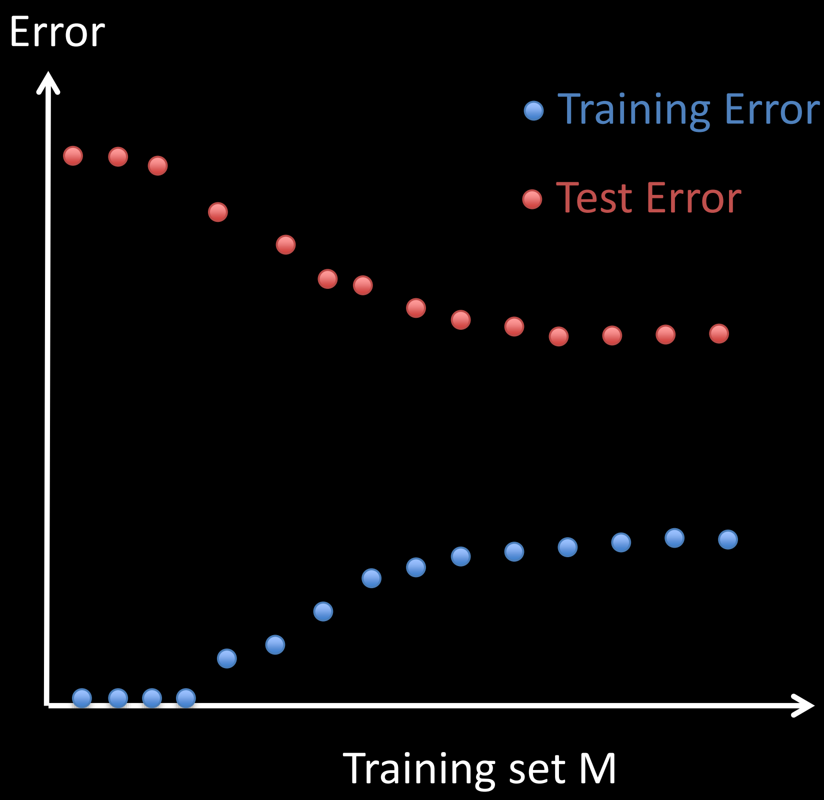
Clicker:

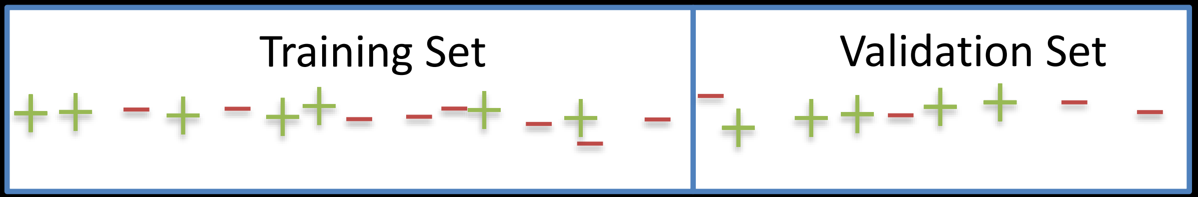
Test error

a) decreases with M
b) increases with M
c) stays constant

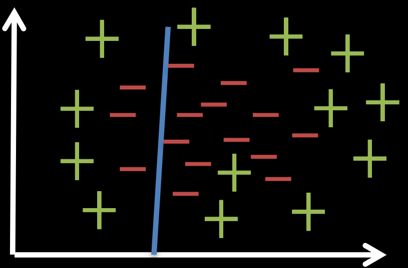


Bias and Variance



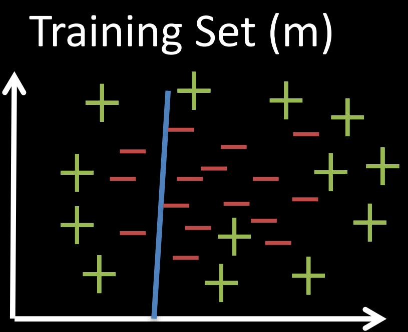


High Bias


Training Set (m)


Error
Test Error
Training Error
Training set M

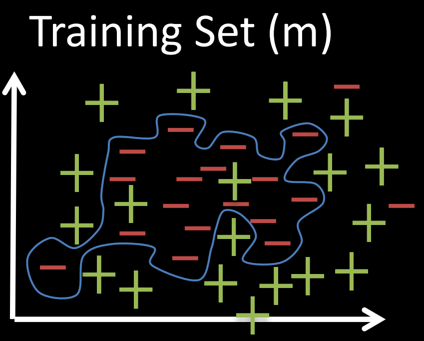
High Bias

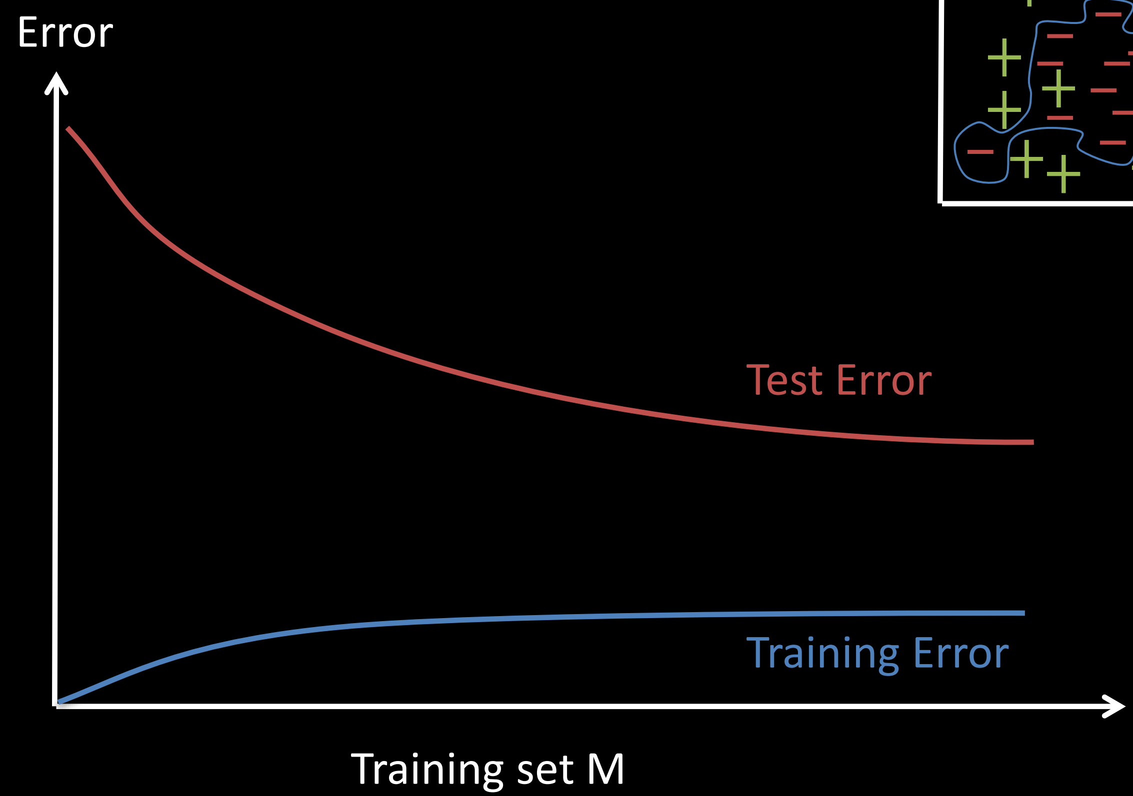


Clicker: If you have high-bias, does more data help?

a) No
b) Yes

High Variance





Clicker: If you have high-variance, does more data help?

a) No
b) Yes

Clicker

1. **Get more training examples**
2. Try smaller sets of features
3. Try getting additional features
4. Try adding polynomial features (kernels)
5. Try increase regularization
6. Try decrease regularization

Helps with

A. High Variance
B. High Bias
C. Both
D. None

Clicker

1. Get more training examples
2. **Try smaller sets of features**
3. Try getting additional features
4. Try adding polynomial features (kernels)
5. Try increase regularization
6. Try decrease regularization

Helps with

A. High Variance
B. High Bias
C. Both
D. None

Clicker

1. Get more training examples
2. Try smaller sets of features
3. **Try getting additional features**
4. Try adding polynomial features (kernels)
5. Try increase regularization
6. Try decrease regularization

Helps with

A. High Variance
B. High Bias
C. Both
D. None

Clicker

1. Get more training examples
2. Try smaller sets of features
3. Try getting additional features
4. **Try adding polynomial features (kernels)**
5. Try increase regularization
6. Try decrease regularization

Helps with

A. High Variance
B. High Bias
C. Both
D. None

Clicker

1. Get more training examples
2. Try smaller sets of features
3. Try getting additional features
4. Try adding polynomial features (kernels)
5. **Try increase regularization**
6. Try decrease regularization

Helps with

A. High Variance
B. High Bias
C. Both
D. None

Clicker

1. Get more training examples
2. Try smaller sets of features
3. Try getting additional features
4. Try adding polynomial features (kernels)
5. Try increase regularization
6. **Try decrease regularization**

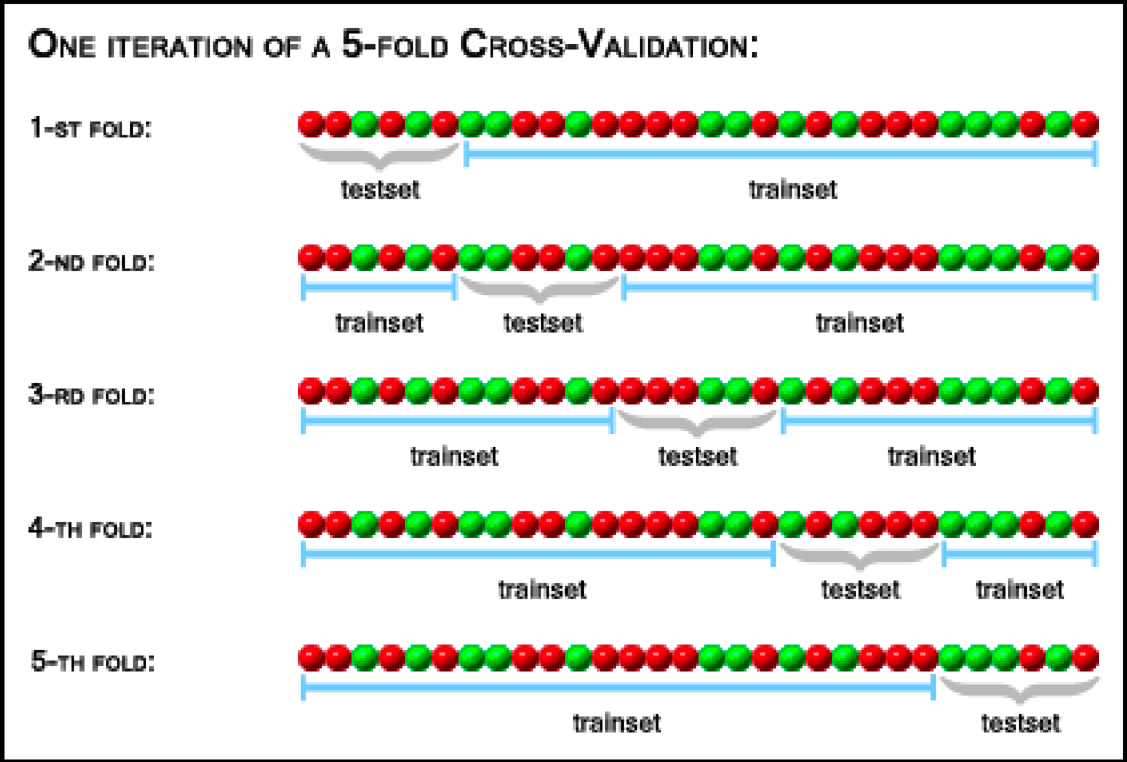
Helps with

A. High Variance
B. High Bias
C. Both
D. None

Cross-validation

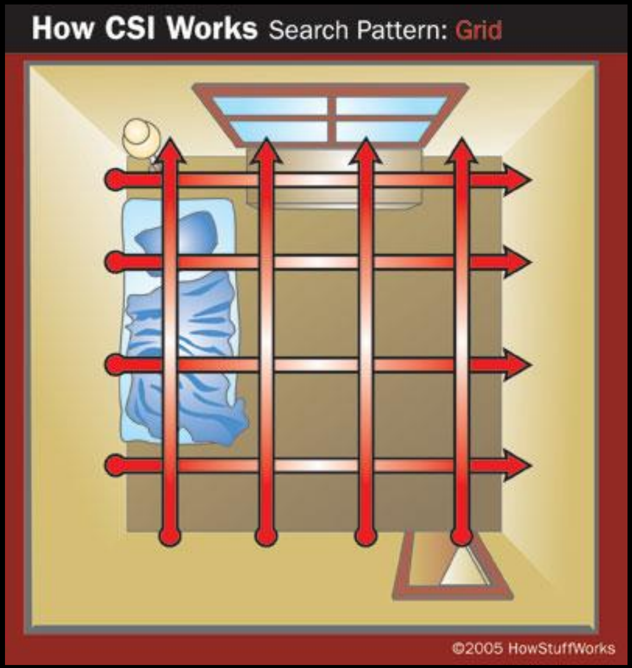
k-fold: split the data into k groups, train on every group except for one, which you test on.

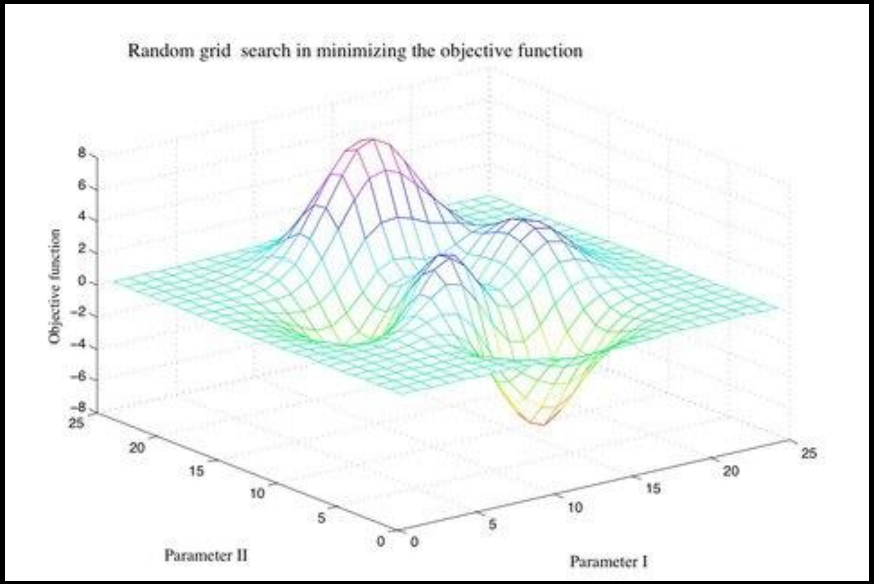
Repeat for all groups



Parameter Tuning

Grid Search





How to speed-up tuning?

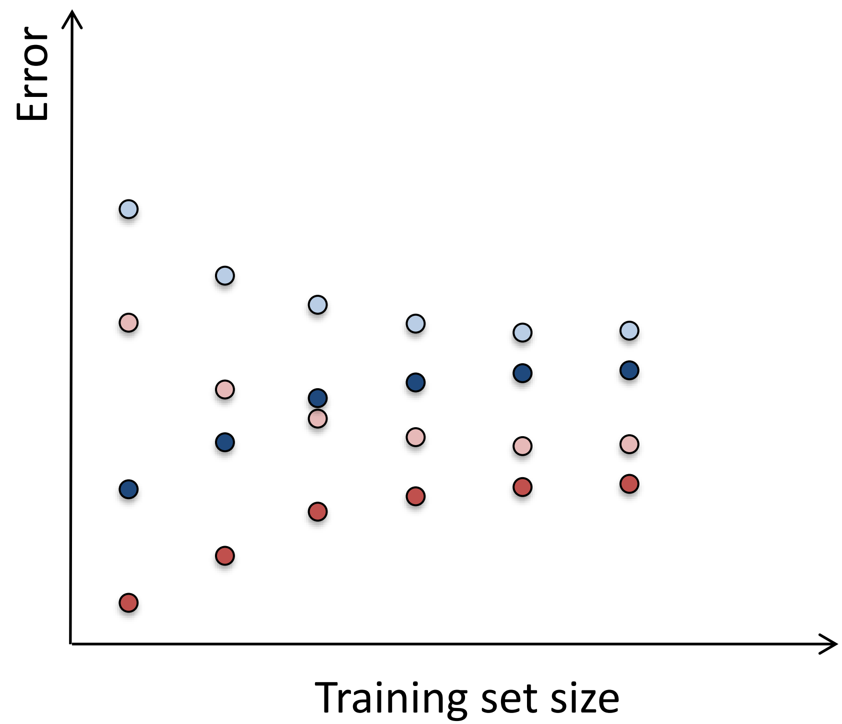
Can we use sampling?

Algorithm 1:

Training Validation

Algorithm 2:

Training Validation



How to speed-up tuning?

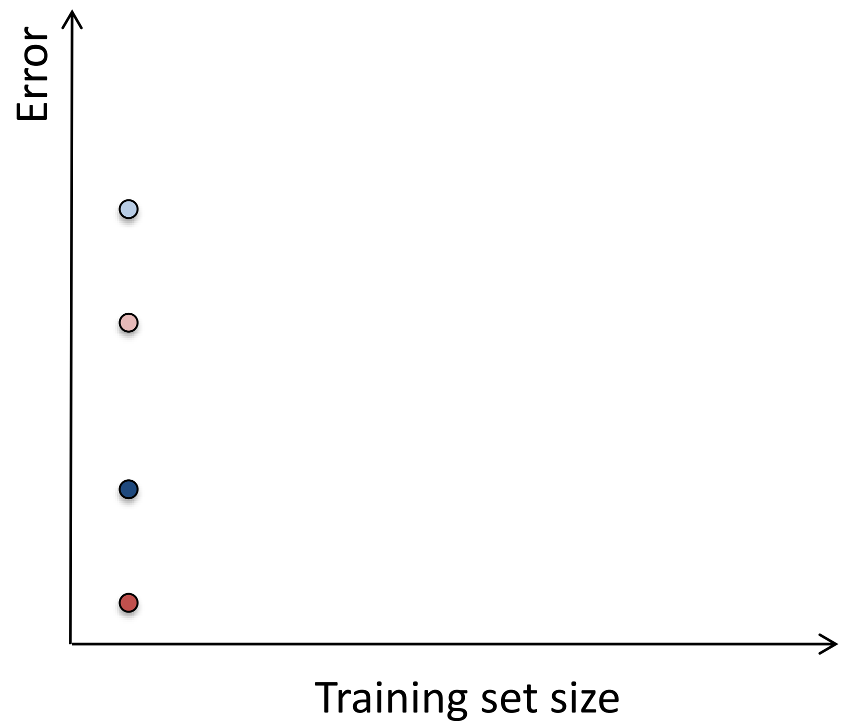
Can we use sampling?

Algorithm 1:

Training Validation

Algorithm 2:

Training Validation



How to speed-up tuning?

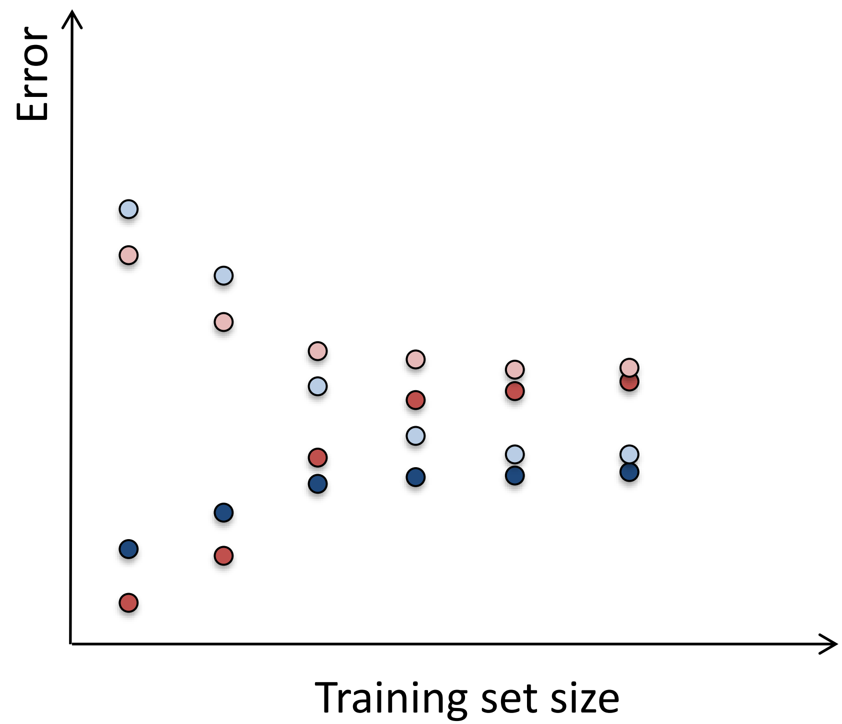
Can we use sampling?

Algorithm 1:

Training Validation

Algorithm 2:

Training Validation



How to speed-up tuning?

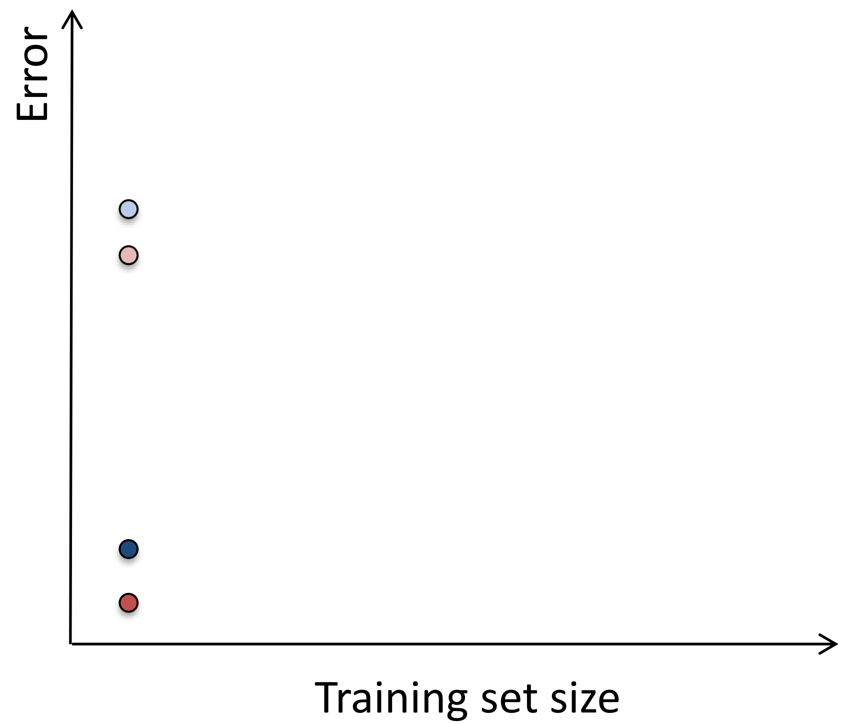
Can we use sampling?

Algorithm 1:

Training   Validation

Algorithm 2:

Training   Validation



How to speed-up tuning?

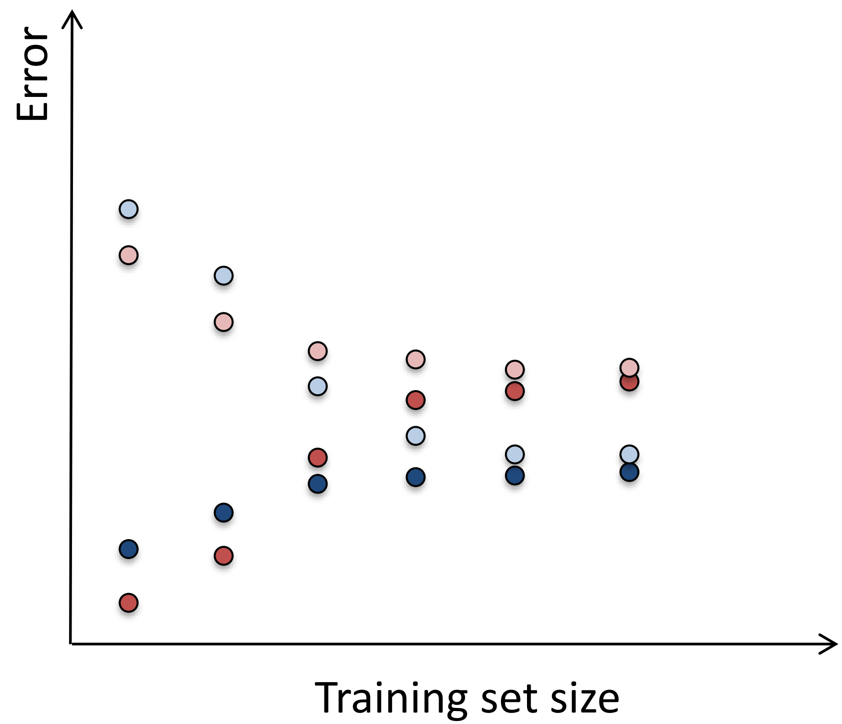
Can we use sampling?

Algorithm 1:

- Training
- Validation

Algorithm 2:

- Training
- Validation



How to speed-up tuning?

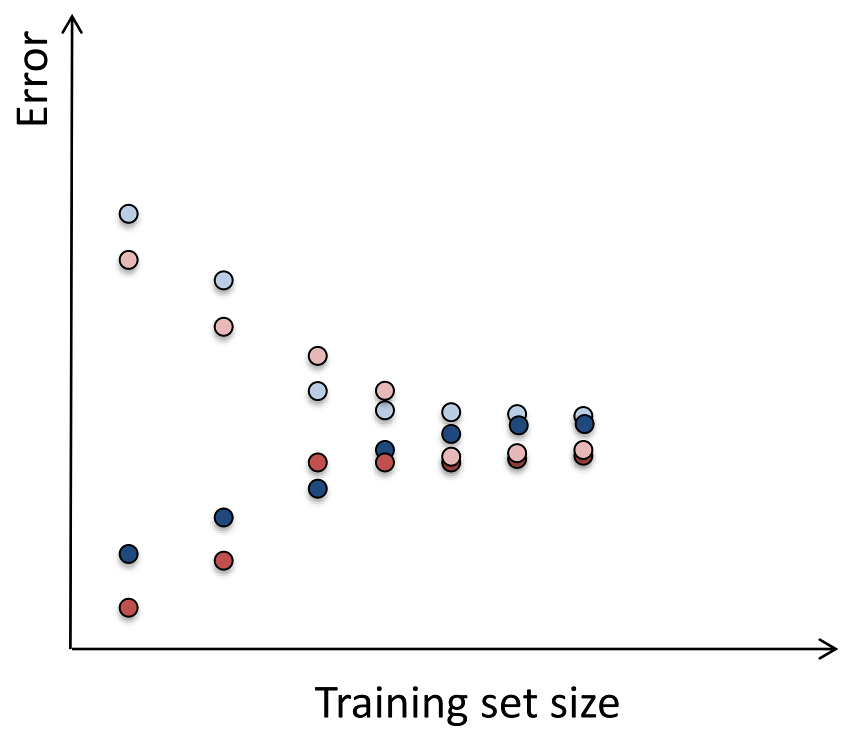
Can we use sampling?

Algorithm 1:

Training Validation

Algorithm 2:

Training Validation



Can we prune now?

How to speed-up tuning?

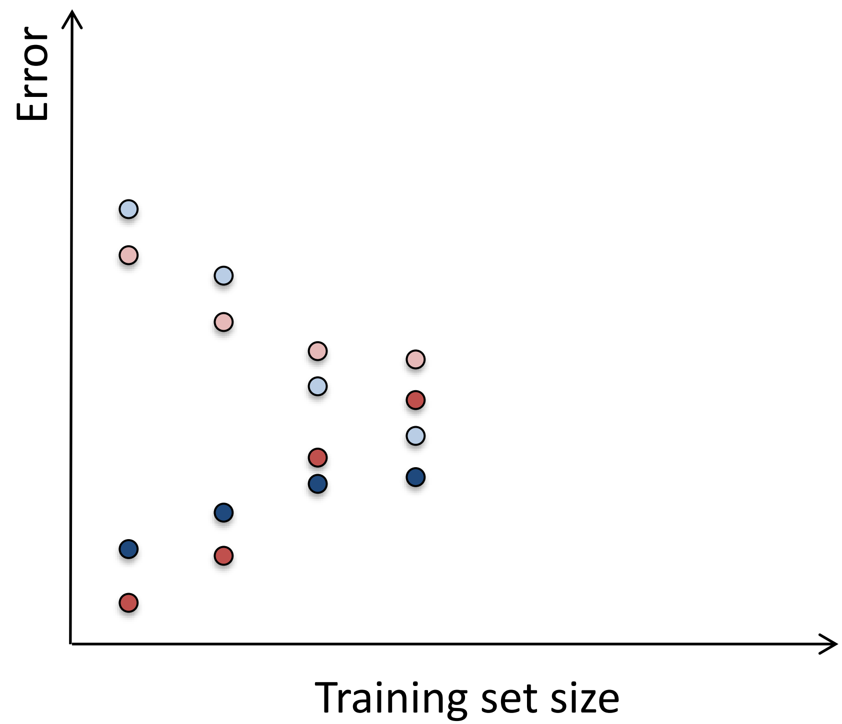
Can we use sampling?

Algorithm 1:

Training   Validation

Algorithm 2:

Training   Validation

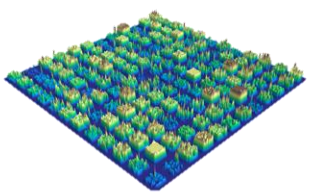


Algorithm 1 training error > Algorithm 2 validation error

Northstar’s (now Einblick) AutoML

Built for *interactive results*,
unlike all other Auto-ML tools, which can take hours to produce results

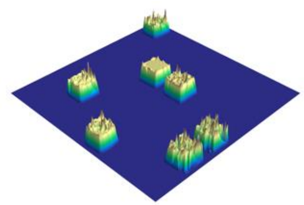
What modeling options do I have?



Rule-based Search Space Expansion

ML pipeline search space visualization

What should I try first?



Use past experience to optimize for expected quality per time-unit

How can I get some quick results?

Adaptive sampling-based pruning and transfer learning

improve based on results

Zeyuan Shang, Emanuel Zgraggen, Benedetto Buratti, Ferdinand Kossmann, Philipp Eichmann, Yeounoh Chung, Carsten Binnig, Eli Upfal, Tim Kraska:
Democratizing Data Science through Interactive Curation of ML Pipelines.
SIGMOD Conference 2019: 1171-1188

https://staging.einblick.ai/?w=6228f27140d515321ee4d967