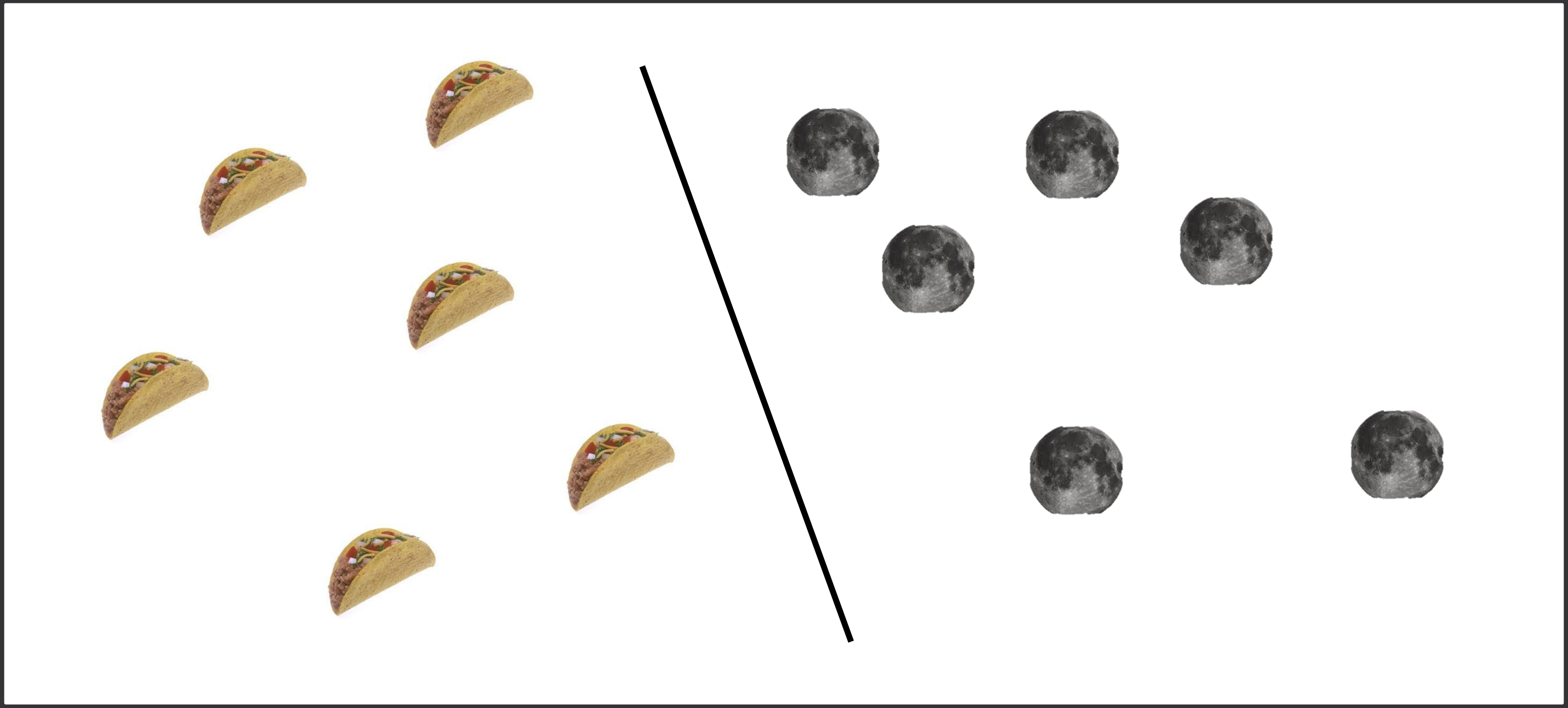
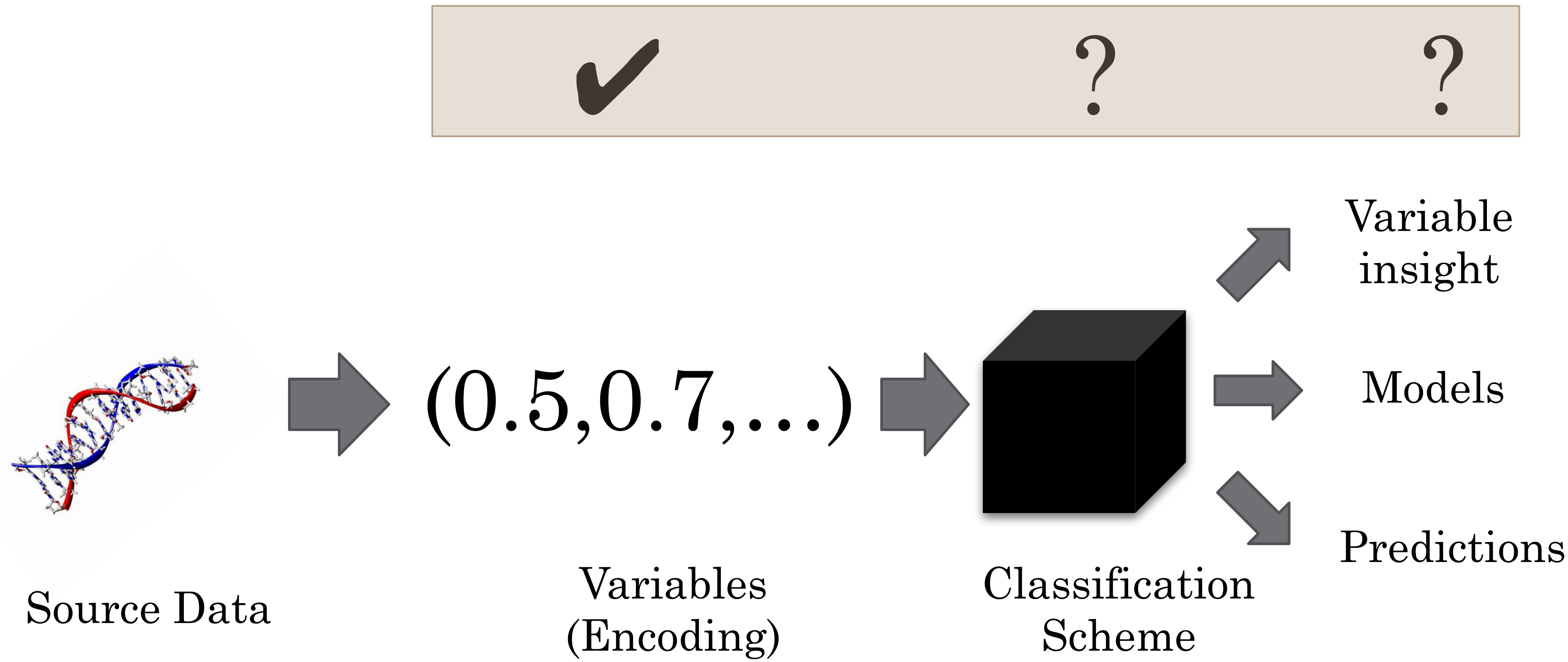


Training a classifier

CSCI 4181 / 6802 Module 1-TRAI

Overview

1. General properties of learning problems
2. Training, testing and quantifying accuracy
3. Choosing a classifier

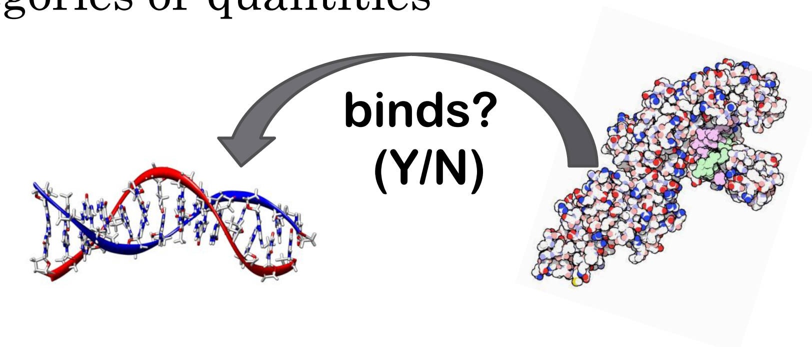

✔
?
?
Source Data
(0.5,0.7,…)
Variables
(Encoding)
Classification
Scheme
Variable
insight
Models
Predictions

Learning Problems

Map input variables to categories or quantities

CLASSIFICATION

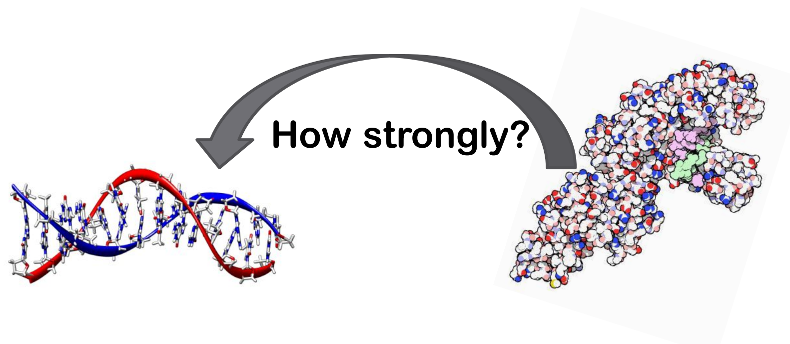
Predict qualitative traits (categories)



REGRESSION

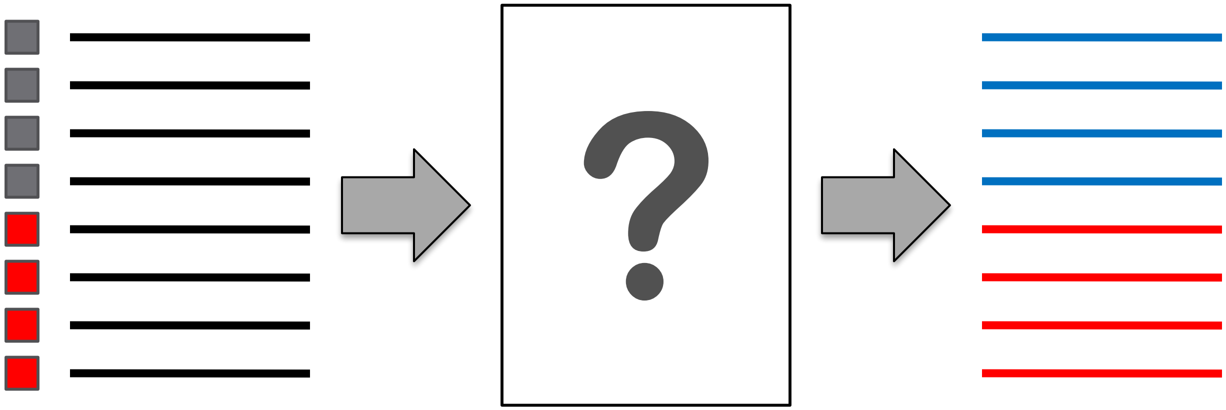
(and related methods)

Quantitative predictions



Training

Goal: to learn the rules (or fit functions) that distinguish classes



What are the properties of a good training set?

- Rnodam spamle from the population
- Sufficiently large
- All classes represented

Types of Learning

SUPERVISED

- Labeled classes
- Feedback: information about labeling is used to train classifier

UNSUPERVISED

- Classes may be labeled or unlabelled
- Classifier develops the classification scheme independently from class labels

SEMI-SUPERVISED

- Use both labeled and unlabeled data
- Unlabeled data can augment knowledge about probability distributions

REINFORCEMENT

- Identify optimal moves through a search space
- Good strategies are rewarded (consider short-term vs long-term tradeoffs)

Goal of supervised learning

Minimize error

(via for instance a *loss function*)

on the training set

e.g., Squared error loss: $\mathrm{EPE}(f) = \mathrm{E}(\mathrm{Y} - f(\mathrm{X}))^2$

Expected prediction error

Expectation

Difference between actual value (Y) and prediction

Hastie section 2.4

Some methods have closed-form solutions that are globally optimal on the cost function

- Many statistical methods e.g. discriminant function analysis, linear regression

Others must use heuristics (iterative training, greedy approaches)

- Neural networks
- Random forests
- Support vector machines

Generalization

A classifier is of little use if it can only do well on data it has been trained on

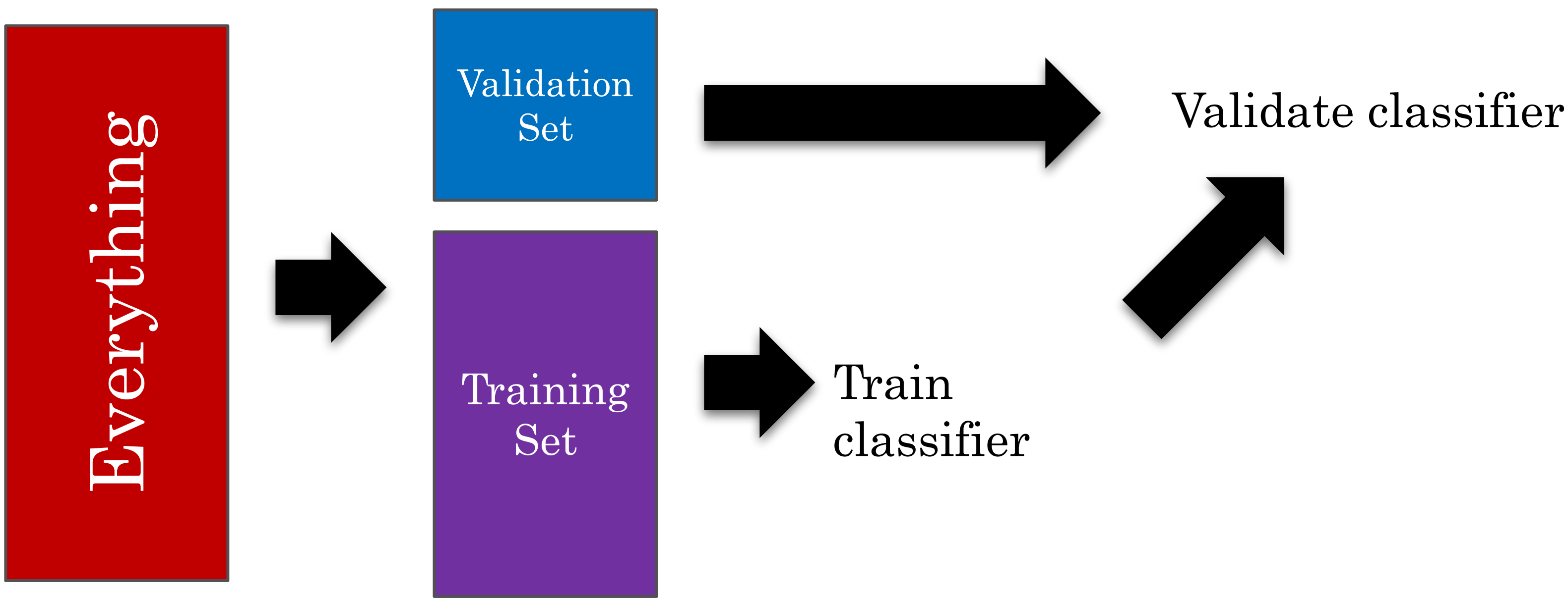
How well does the classifier handle cases that were not present in the training set?

General form

Training data → Validation data → Test data

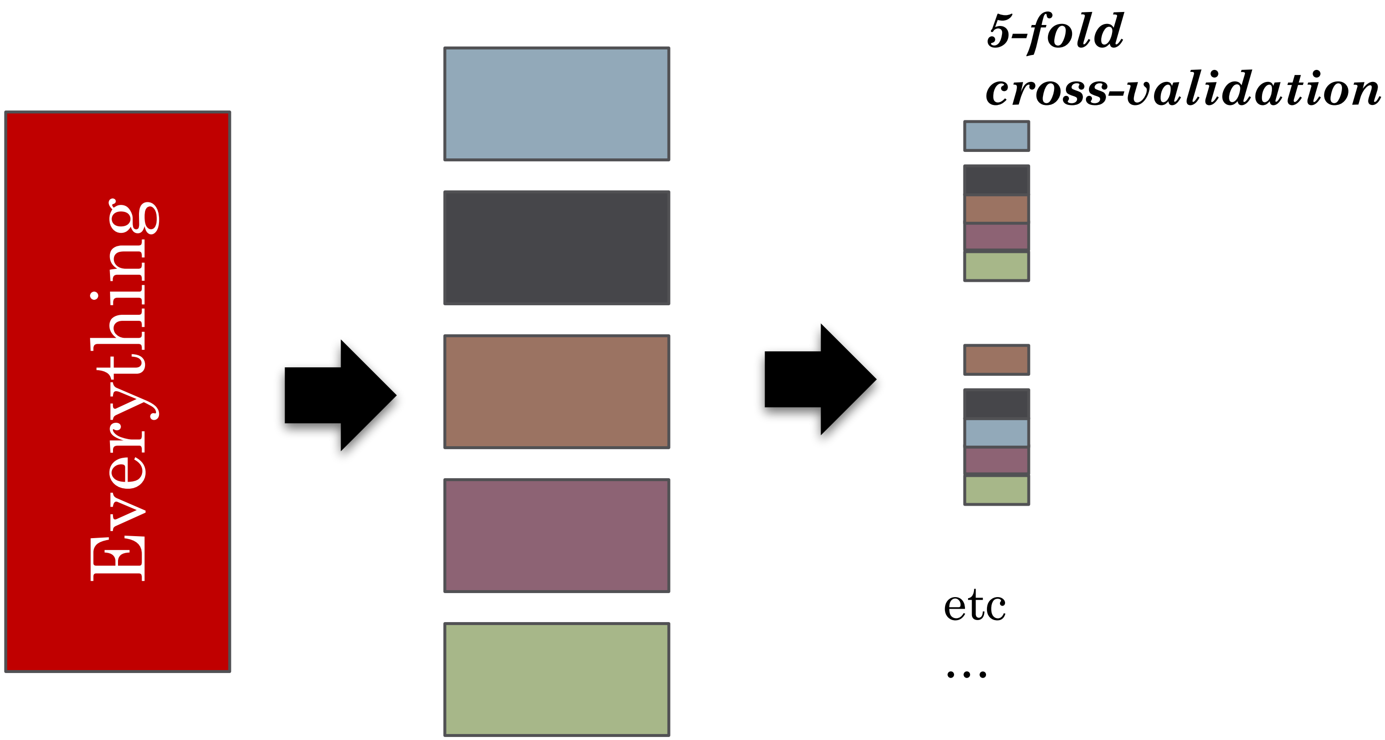
Data set splitting
(*holdout* method)

Use a fraction of available cases as the *training set*, reserve the remainder for a *validation* set



Cross-validation

Repeated training with different subsets



The *cross-validation score* is the average performance on all validation sets

Sample sets at random, but make sure every class is represented!

In the two-class case:

+ training set

- training set

+ validation set

- validation set

Classification Accuracy

CONFUSION MATRIX for a two-class (positive and negative set) problem

Predicted / True	+	-
+	True positive	False negative
-	False positive	True negative

may require THRESHOLDING of continuous predictions

Quantifying Accuracy

Sensitivity, recall, true positive rate	$= \frac{TP}{TP+FN} = \frac{TP}{n^+}$	Specificity, true negative rate	$= \frac{TN}{TN+FP} = \frac{TN}{n^-}$
Positive predictive value, precision	$= \frac{TP}{TP+FP}$	Negative predictive value	$= \frac{TN}{TN+FN}$
False positive rate, fallout	$= \frac{FP}{FP+TN} = \frac{FP}{n^-}$	False discovery rate	$= \frac{FP}{FP+TP}$

Don't forget that regularization may impact scoring!

Chicco and Jurman (2020) *BMC Bioinformatics*

Matthews Correlation Coefficient

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

F_1 Score

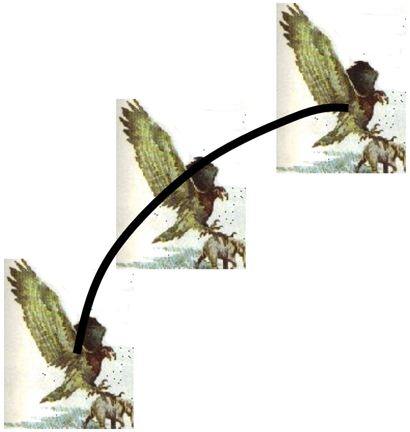
$$F_1 = \frac{2}{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} = \frac{\mathrm{tp}}{\mathrm{tp} + \frac{1}{2}(\mathrm{fp} + \mathrm{fn})}$$

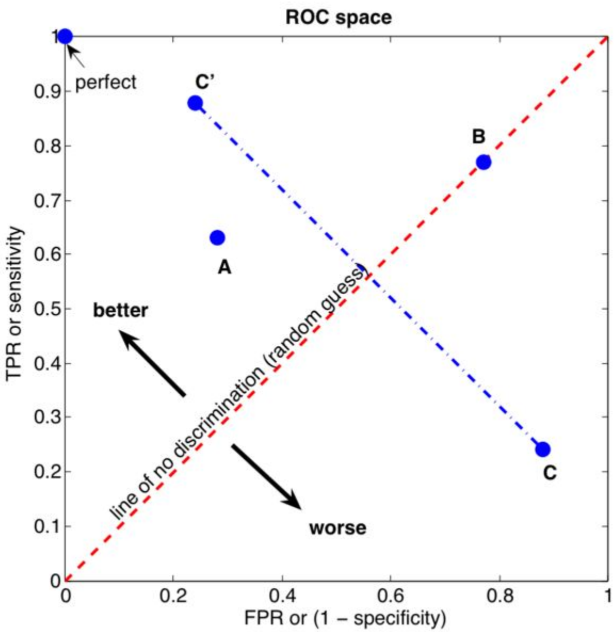
'Balanced' accuracy:

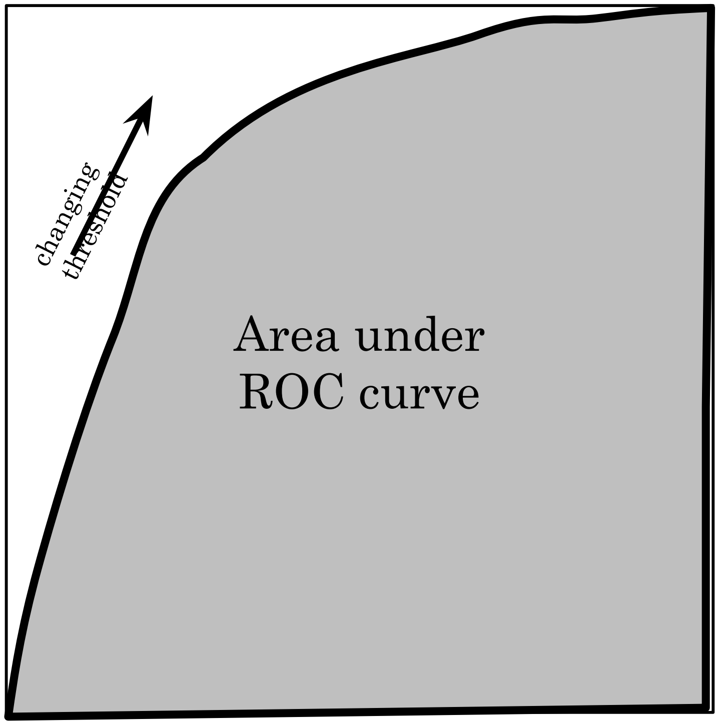
[**TP / (TP + FN)** + **TN / (TN + FP)**] / 2

Others: see Baldi et al. (2000) in *Bioinformatics*

ROC Curves







(Wikimedia Commons)

Regression problems

Mean absolute error

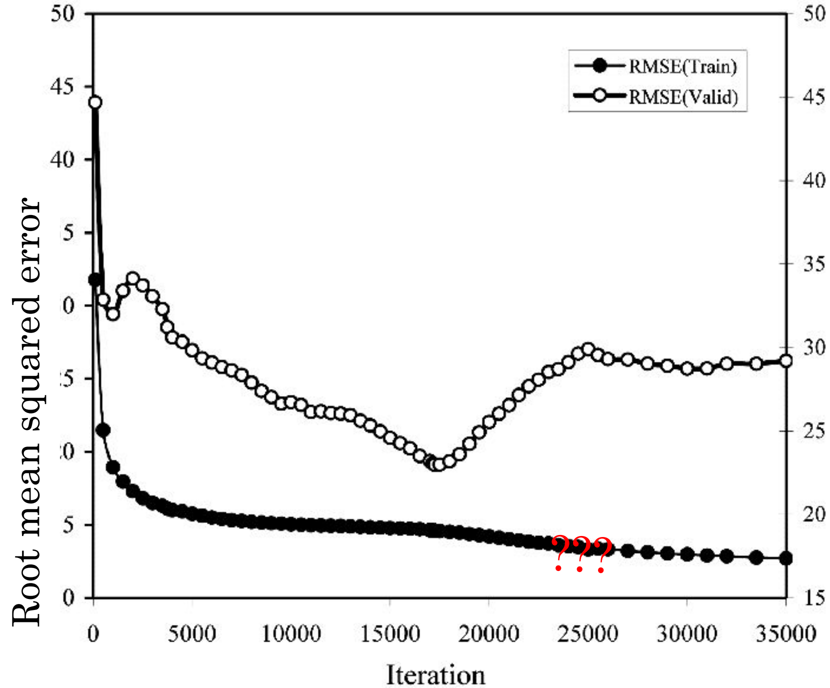
$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left| y_i - x_i \right|}{n} = \frac{\sum_{i=1}^{n} \left| e_i \right|}{n}.$$

Difference between predicted and true values

Root mean square deviation

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{t=1}^{T} (\hat{y}_t - y_t)^2}{T}}.$$

Iterative training



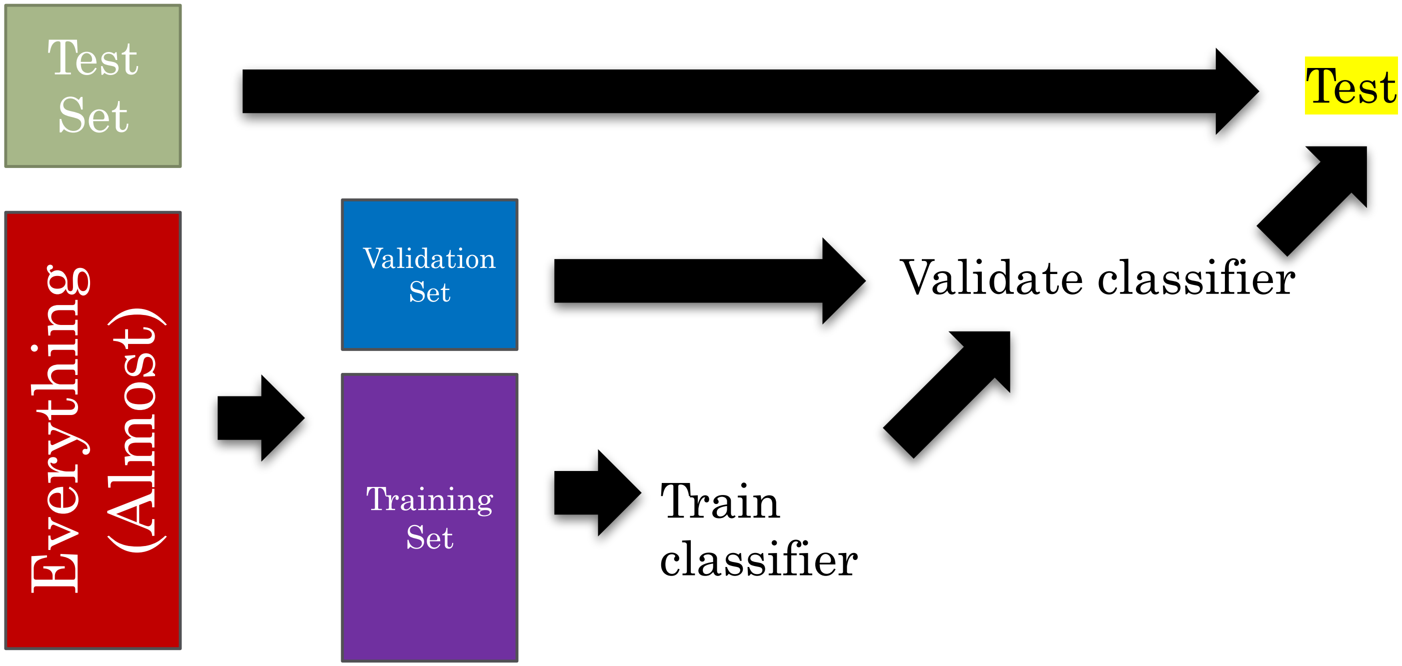
Monotonic!

Training set accuracy improves, but at some point
validation set accuracy may go boom
= OVERFITTING

Habibi-Yangjeh, Aziz, and Mahdi Esmailian. (2007) *Bull Kor Chem Soc*

Test set

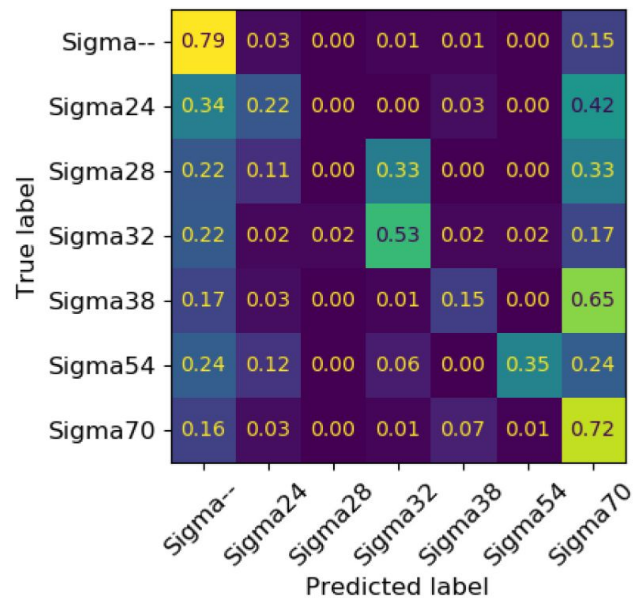
- With k-fold cross-validation, ALL data are used in training k-1 times
- So it is common practice to hold out part of the data entirely and assess only AFTER cross-validation / parameter selection has been completed



Expanding to multiple classes

- There's nothing special about "Positive" and "Negative" classes – invert the labels and corresponding scores would either change or map deterministically
- Do we weight all classes equally, or do we weight by abundance?

Promoter predictions by class – different promoter types are active at different times



Rafante and Beiko (submitted)

Key questions

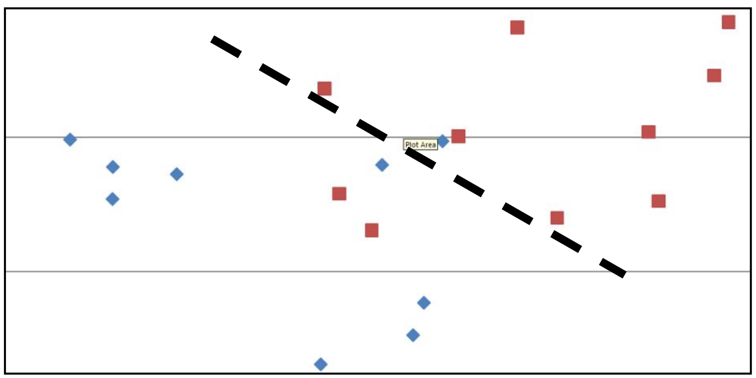
1. Which is more desirable, sensitivity or specificity?
2. How many folds of cross-validation is the right number of folds of cross-validation?
3. What is the value of our classifier if the accuracy on the test set is 60%?

No one classifier is best for every classification problem

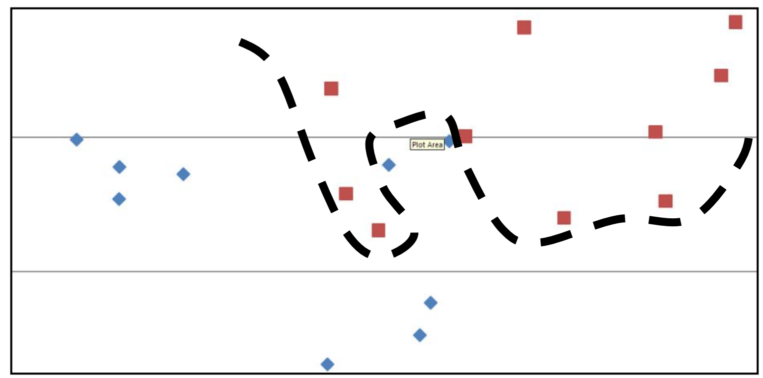
What criteria should we consider?

Bias-Variance Tradeoff

Do we want a classifier that is as simple as possible, or one that can make complex decisions?



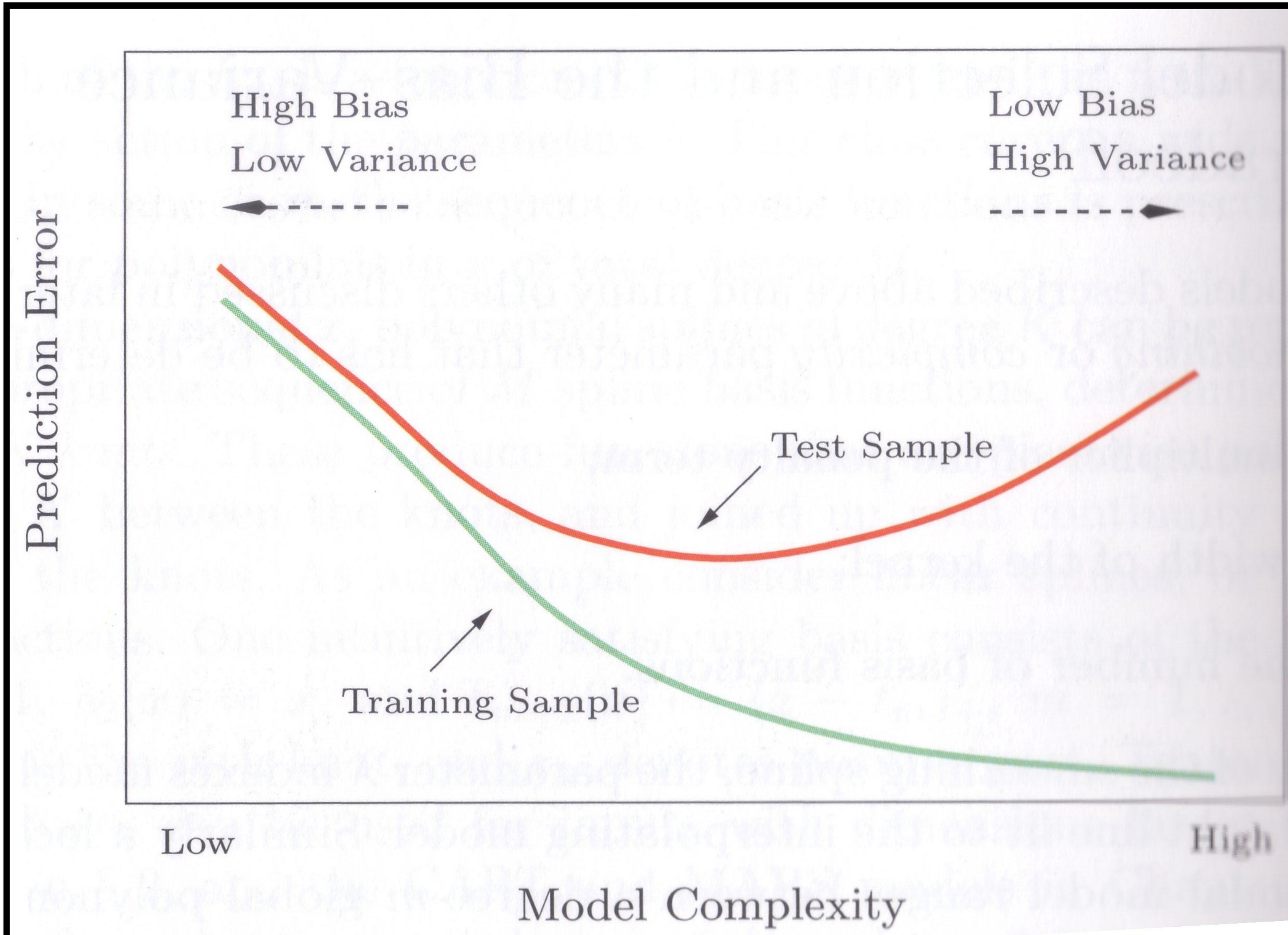
Bi
as



Variance
(overfitting)

the ability of the machine to learn any training set without error. A machine with too much capacity is like a botanist with a photographic memory who, when presented with a new tree, concludes that it is not a tree because it has a different number of leaves from anything she has seen before; a machine with too little capacity is like the botanist's lazy brother, who declares that if it's green, it's a tree. Neither can generalize well. The exploration and

Burges 1997, “A Tutorial on Support Vector Machines for Pattern Recognition”.

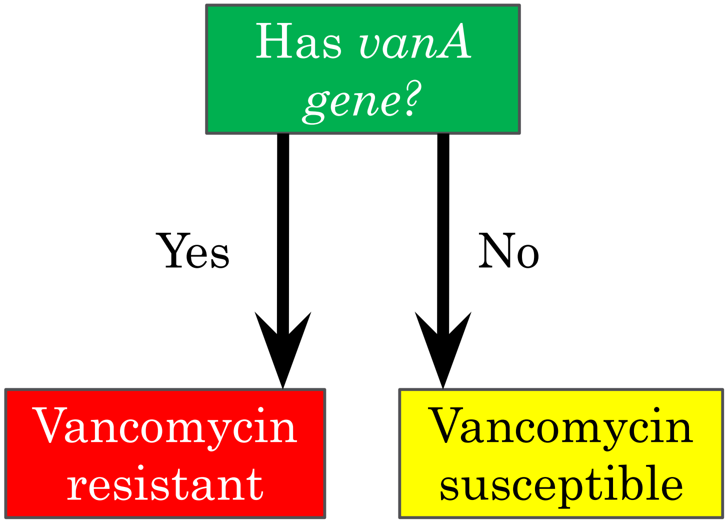

High Bias
Low Variance
Low Bias
High Variance
Prediction Error
Test Sample
Training Sample
Low
High
Model Complexity

Hastie, p.38

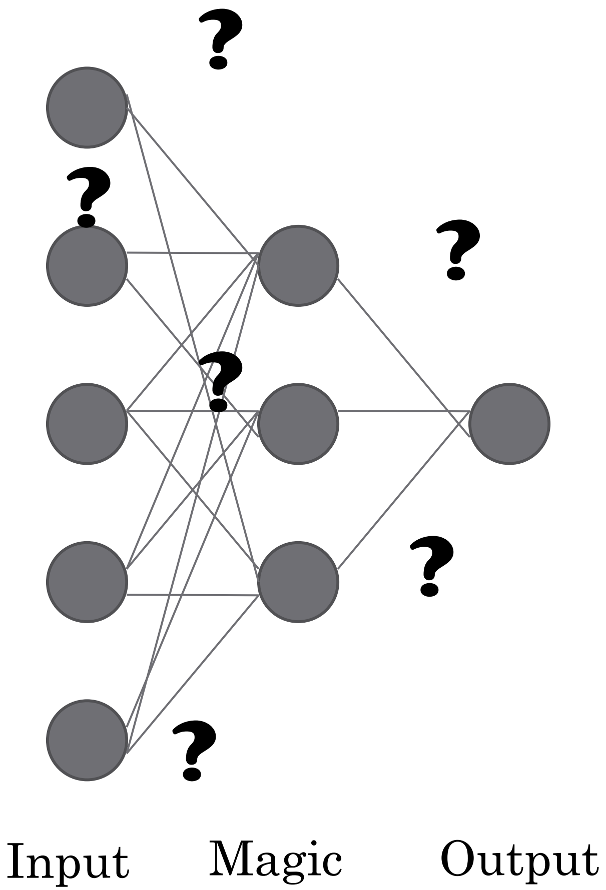
Interpretability

Some methods yield understandable (or almost understandable) rules, others do not

Decision tree



Artificial neural network



Tractability

If the training data are necessarily high-dimensional, then a simpler classifier may be necessary

(or we need to be more aggressive in our feature selection / extraction)