GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

LECTURE 14: CLASSIFICATION REVIEW

I. SUPPORT VECTOR MACHINES
II. REGULARIZATION
III. KERNELS

Questions?

DATA EXPLORATION

SUPERVISED LEARNING: REGRESSION

SUPERVISED LEARNING: CLASSIFICATION

- LOGISTIC REGRESSION
- NAIVE BAYES
- RANDOM FORESTS
- SUPPORT VECTOR MACHINES
- COMPETITION (TODAY)

UNSUPERVISED LEARNING

VARIOUS TOPICS

Questions?

I. REVIEW
II. COMPETITION
III. GUEST SPEAKER

Questions?

- APPLYING SUPERVISED LEARNING TECHNIQUES TO A REAL-LIFE PROBLEM

I. REVIEW

	continuous	categorical
supervised	regression	classification
unsupervised	dimension reduction	clustering

What supervised algorithm should I pick for which problem?

Try them all with varying regularization parameters and pick the one with the **best cross-validation results**

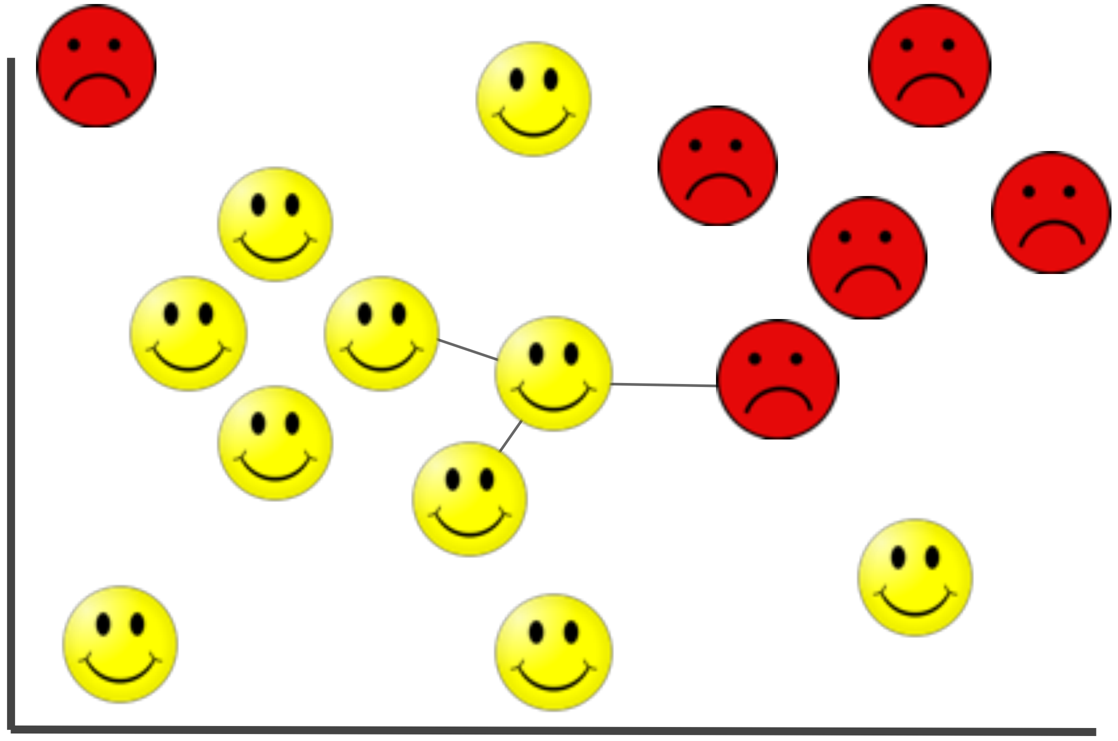
To avoid overfitting on the test set, you might want to use three different sets: training set, cross-validation set, test set

	kNN	*Logistic*	*NB*	*RF*	*SVM*
Linear	–	+	+	–	–
Interpretation	–	+	+	–	–
Feature impact	–	+	+	+	–
Configuration	+	+	+	+	–
Overfitting	*k*	*L1/L2*	Prior	*n* trees	*C, γ, d*
Scalable	–	+	+	–	+/–

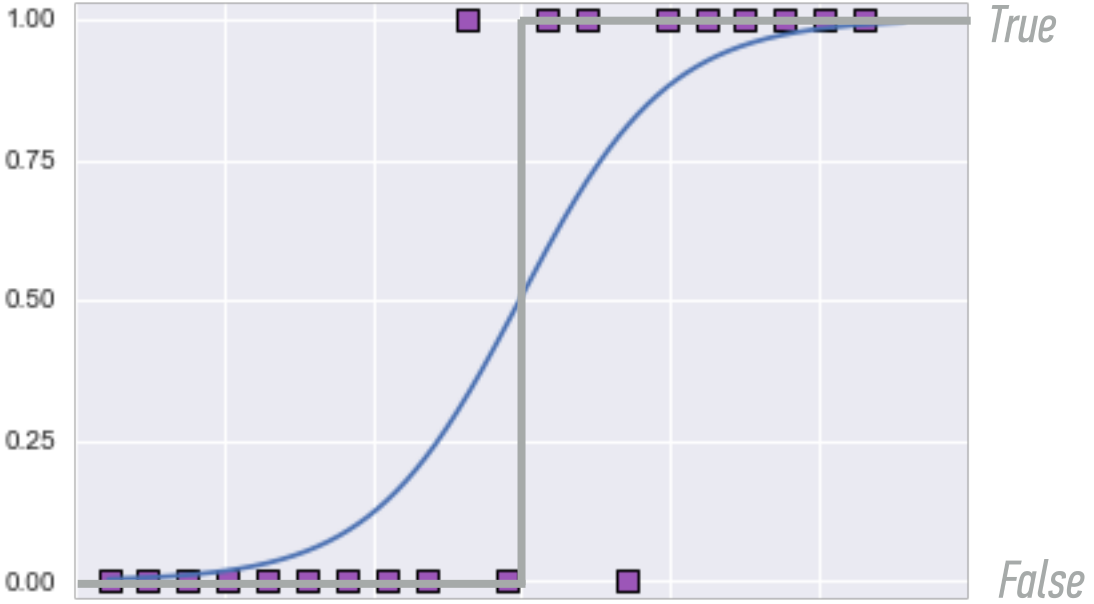
Choose k
e.g., k = 3

Find k nearest neighbors

Take majority vote



Logistic regression gives us predicted probabilities, which then could be 'snapped' to class labels

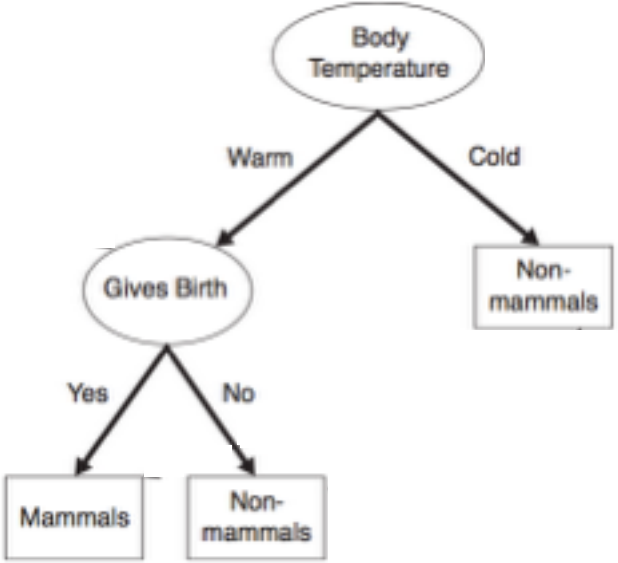


The Naive Bayes algorithm combines the probability of a class C overall with the probabilities of each individual feature appearing in class C

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

$$P(C|\{x_i\}) \sim P(C) \prod_i P(x_i|C)$$

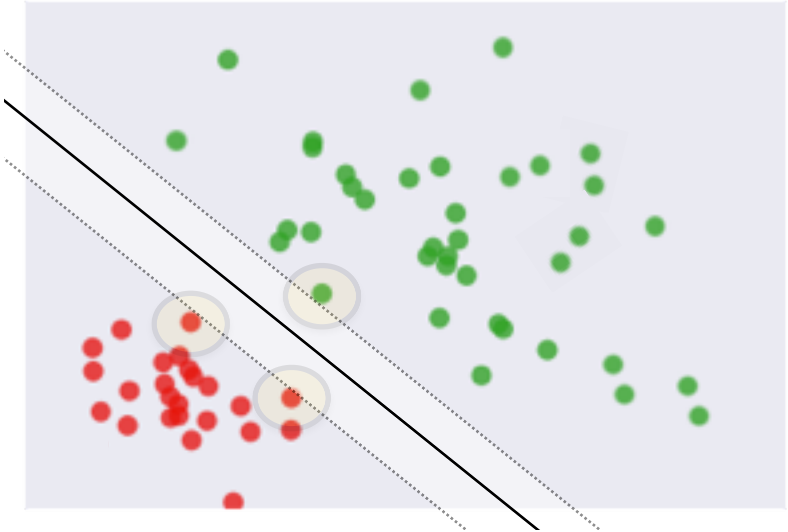
A decision tree for mammal classification…



…may be an accurate way of describing the dataset

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has Legs	Hiber-nates	Class Label
human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

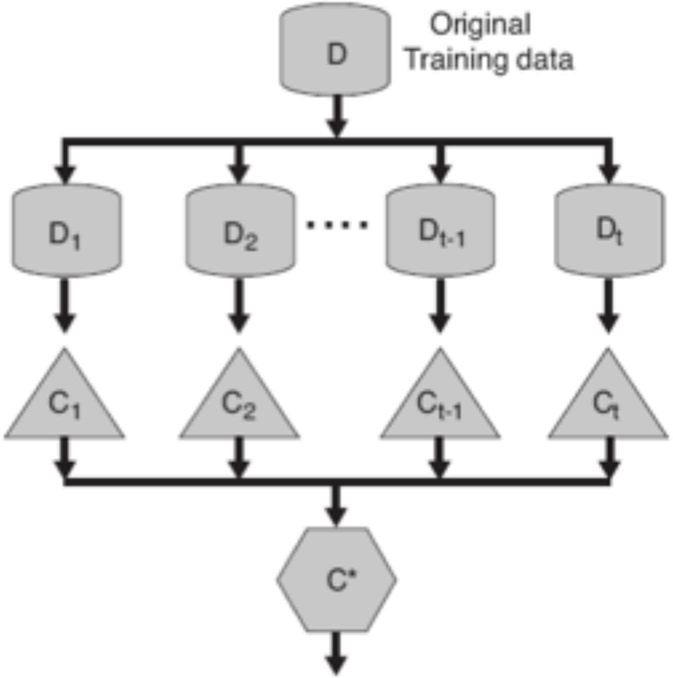
source: http://www-users.cs.umn.edu/~kumar/dmbook/ch4.pdf



Notice that the margin depends only on a subset of the training data – the points nearest to the decision boundary.

These points are called the **support vectors***.*

The other points don't affect the construction of the hyperplane at all!

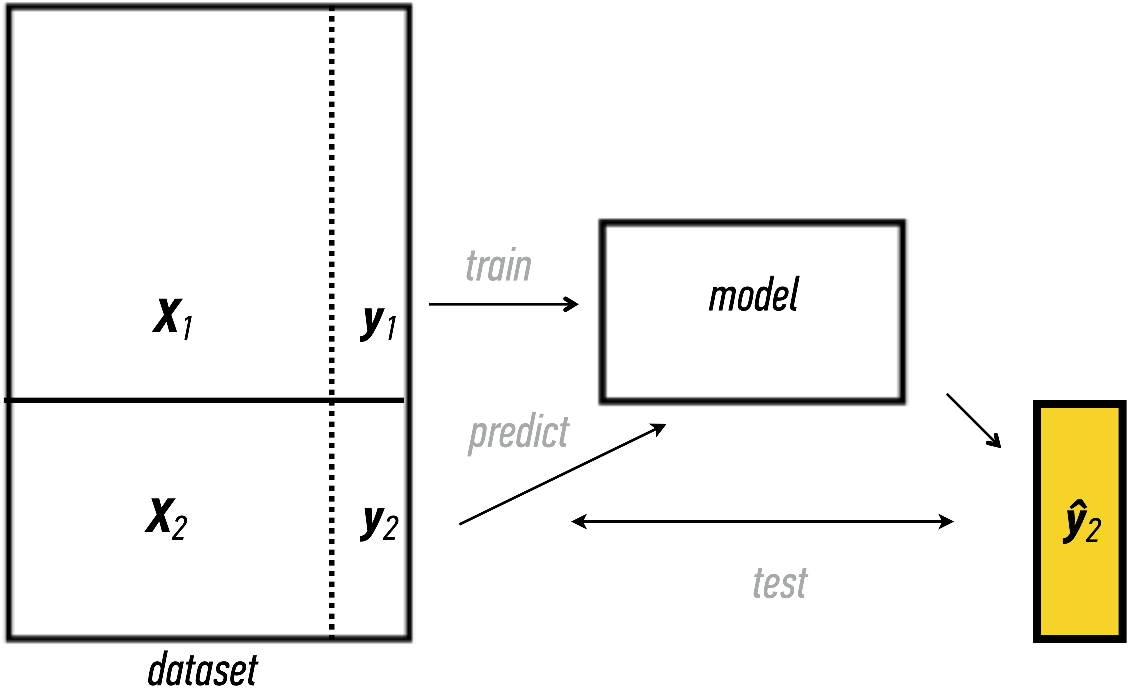


Train your base classifier on different bootstrap samples of your training set and take aggregate vote

Ensemble technique reduce the variance (overfitting), not the bias (underfitting)

How do we test the model's predictions?

*Train model on a part of **X**, and test the results on the rest of the data*



How do we test the model's predictions?

Accuracy = $(TP + TN) / all$

Precision = $TP / (TP + FP)$

% correct of all positive predictions

Recall = $TP / (TP + FN)$

% correct of all positive cases

F1 score = $2 \frac{P \times R}{P + R}$

AUC = % probability a positive case is scored higher than a negative case

	predictions	
truth	**Yes**	**No**
Yes	TP	FN
No	FP	TN

- *When working with* **distance***, scale your features*
  kNN, SVM (MinMaxScaler, StandardScaler)

- *Be wary of* **local minima**
  Decision Trees, non-convex cost functions

- *Be wary of* **bias/variance** *(underfitting/overfitting)*
  Little data, many features ⟶ Overfitting (too complex model)
  Lots of data, few features ⟶ Underfitting (too simple model)

DISCUSSION