Math 407

Support Vector Machines

Unsupervised Learning

Project Instructions

- Choose something you'd like to predict.
- Find a dataset.
- Decide which methods from the class are appropriate to try
- Train models, select the best one
- Estimate your model's performance.
- Write a short paper (details for the format on website)

Project Instructions

Your project should meet at least one of the following criteria:

- Dataset needed extensive cleaning and formatting.
- Tried a method not covered in class.
- Created a prediction model that will be useful for you or someone else

You may work in teams of 1-3 students.

Due date: Friday of Finals week

Support Vector Machine

Tries to find the margin that “best” separates the two categories

- Choice of C, the “budget”
- Choice of kernel
  - Linear
  - Polynomial, also need to choose degree
  - Radial, also need to choose “gamma”

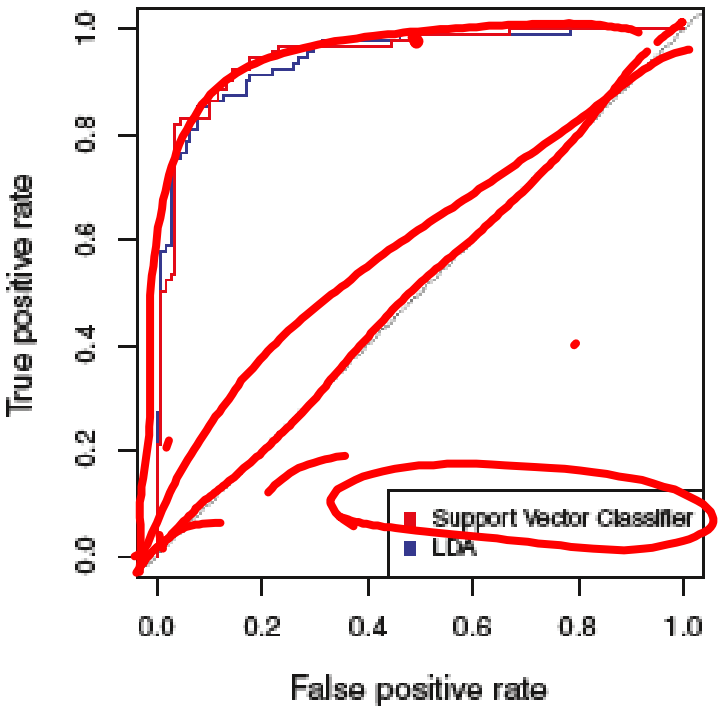
ROC Curves

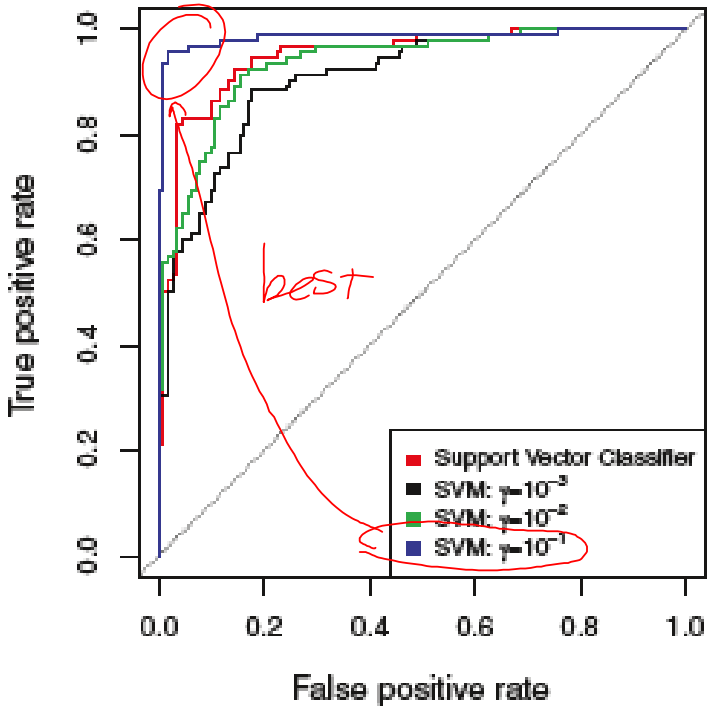
Useful for visualizing how the misclassification rates **by type change** as some "tuning parameter" is changed, such as the degree of a polynomial kernel, gamma for a radial kernel, etc.

Some jargon for types of (mis)classification rates:

- True positive = "sensitivity"
- False positive = "1-specificity"

Example: chose the tuning parameter (in this case, gamma) that achieves a good balance between the types of (mis)classification rates

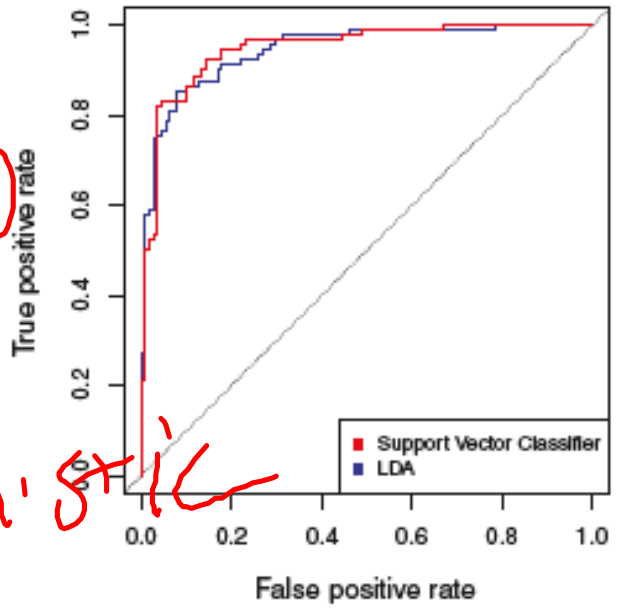


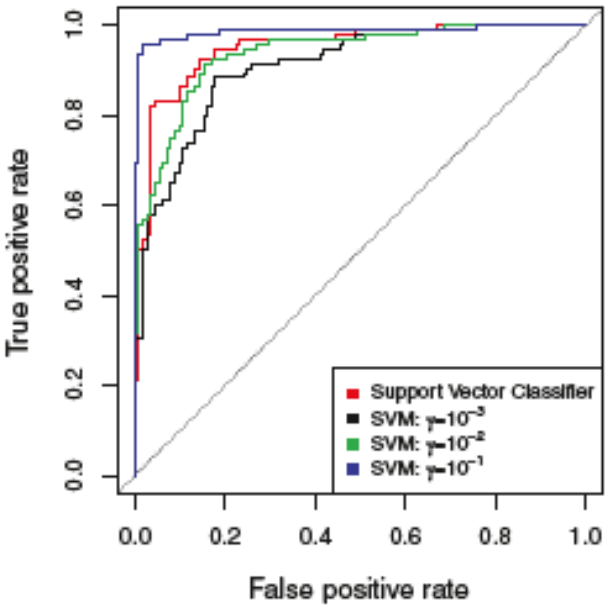


ROC curves should be computed from a test set or averaged over test folds from cross-validation.

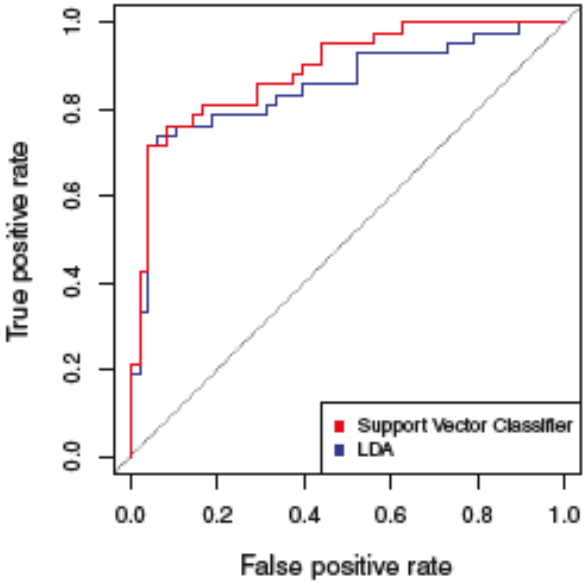
Many functions in R that make ROC curves will by default use the training set and thus show overly optimistic misclassification rates.

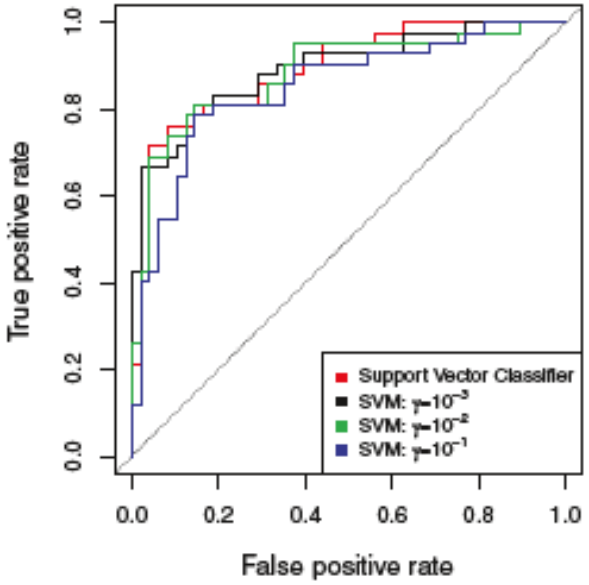
train = too optimistic





test





Methods for classifying a binary outcome Y

- Logistic regression (glm)
- Penalized logistic regression (glmnet)
- Logistic regression with penalized splines (gam)
- LDA, QDA
- SVM
- kNN

Extending methods of binary classification to k-category classification

What if Y is a qualitative variable with K>2 categories?

Two main approaches using a binary classifier:

- One vs. one
- One vs. all

One vs. one

- Train a binary classifier for each pair of categories
- Count how many times a test data point is classified as each possible category, majority vote wins.

- Requires K chose 2 models trained

$$\binom{k}{2} = \frac{k!}{2!(k-2)!}$$

One vs. All

- Train binary classifiers with the response variable of being in a particular category or not
- Classify a test point as the most likely category (or furthest from the decision boundary)
- Requires K models trained

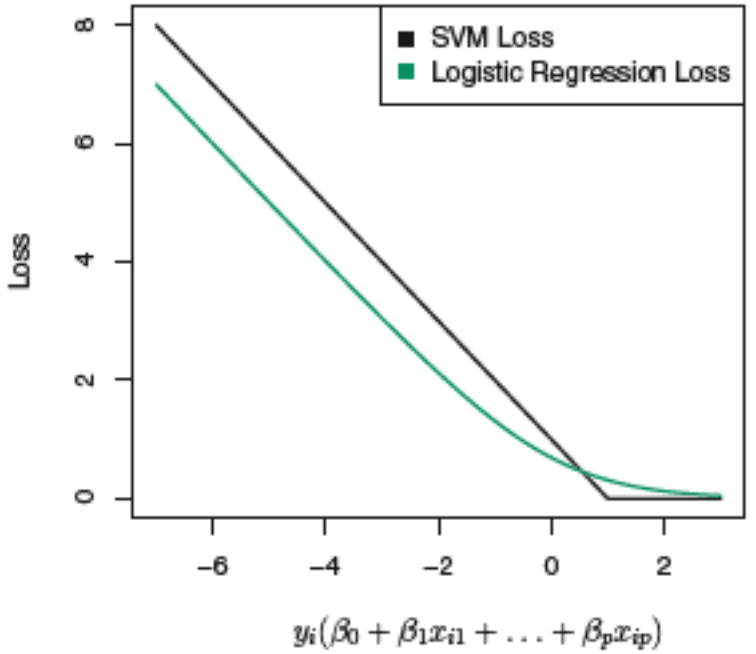
Multinomial regression is an example – models the log odds of being in the kth category vs. not.

Misc. Connection

The methods of logistic regression and SVM both try to optimize an expression of the form

LOSS + λ penalty

Both use the ridge penalty, and the "loss" functions are very similar



Unsupervised Learning

What if we want to classify or predict a response variable Y but only have examples of predictors X in the training set?

Example: gradescope is a software that can be used to classify student work (i.e. responses to an exam question) into categories so that it can be consistently graded.

Unsupervised Learning

Example: sort online customers into categories by their browsing histories – show different ads to different categories of customers

Example: sort online customers into categories by their browsing histories and locations – assign them different prices to pay for items

Popular methods of unsupervised learning

- PCA/SVD
- MDS
- K means clustering
- Hierarchical clustering / Dendrogram

K means clustering

- Choose the number of clusters, K
- Divide the dataset into K clusters so that the **variation within a cluster** is minimized

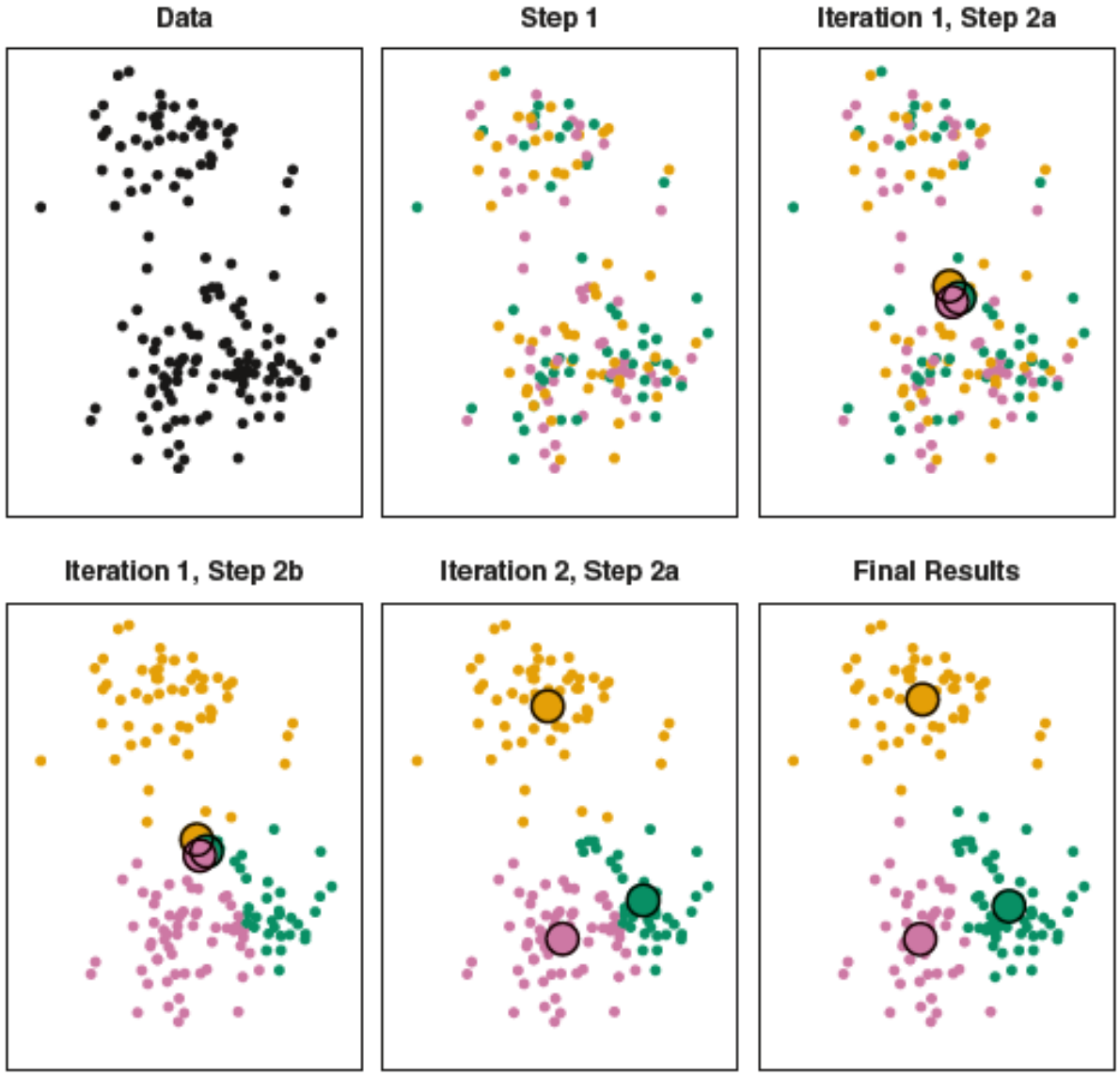
K means clustering

- Choose the number of clusters, K
- Divide the dataset into K clusters so that the **variation within a cluster** is minimized

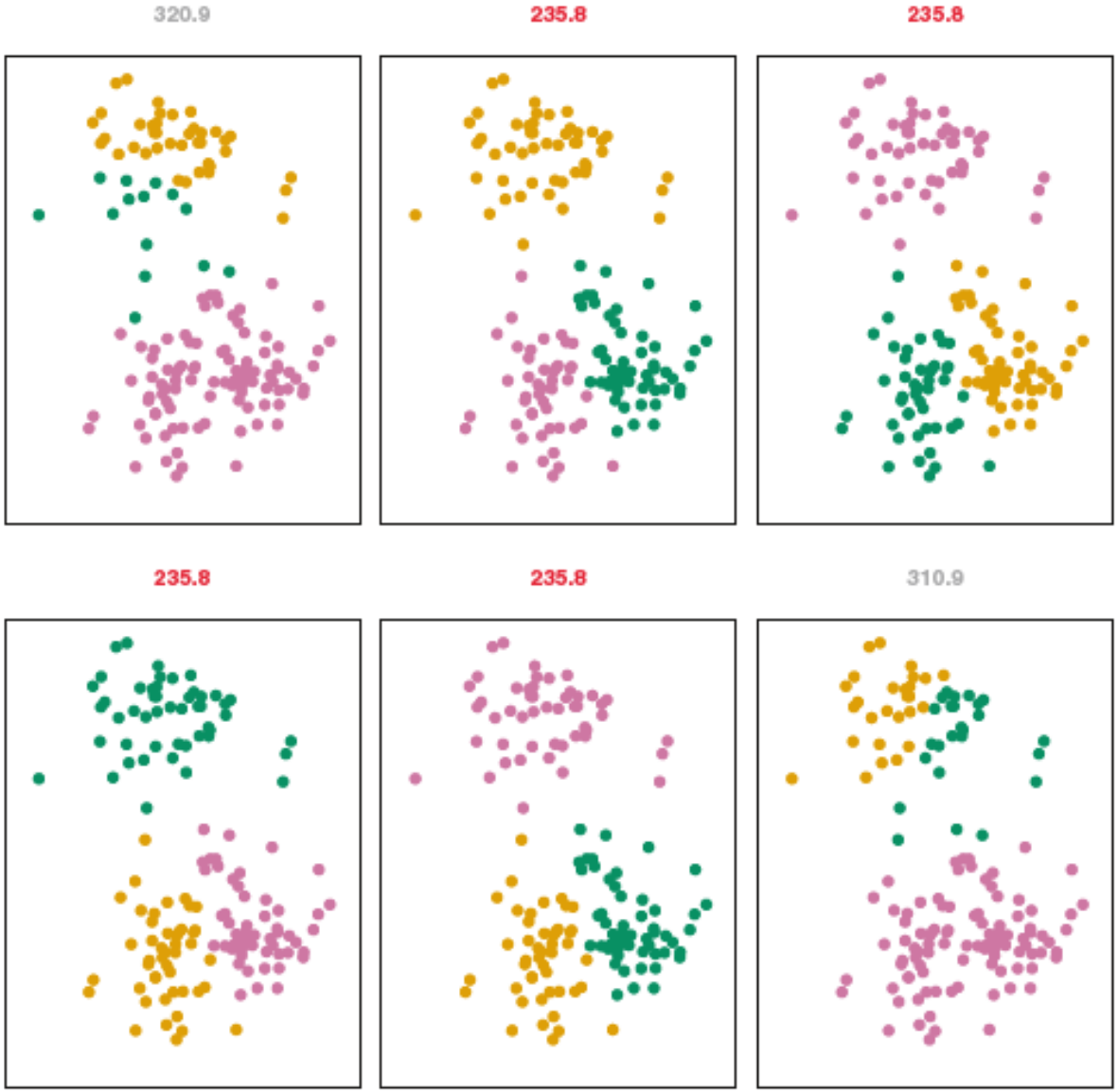
Note that no estimate of the misclassification rate can be had because we have no Y values…

…so there's nothing to cross-validate or test

Algorithm: assumes Euclidean distance to measure “closeness”



Sometimes a local min is found, not the global min



Example: colors are the result of k-means