LECTURE 20:
SUPPORT VECTOR MACHINES PT. 1

November 27, 2017
SDS 293: Machine Learning

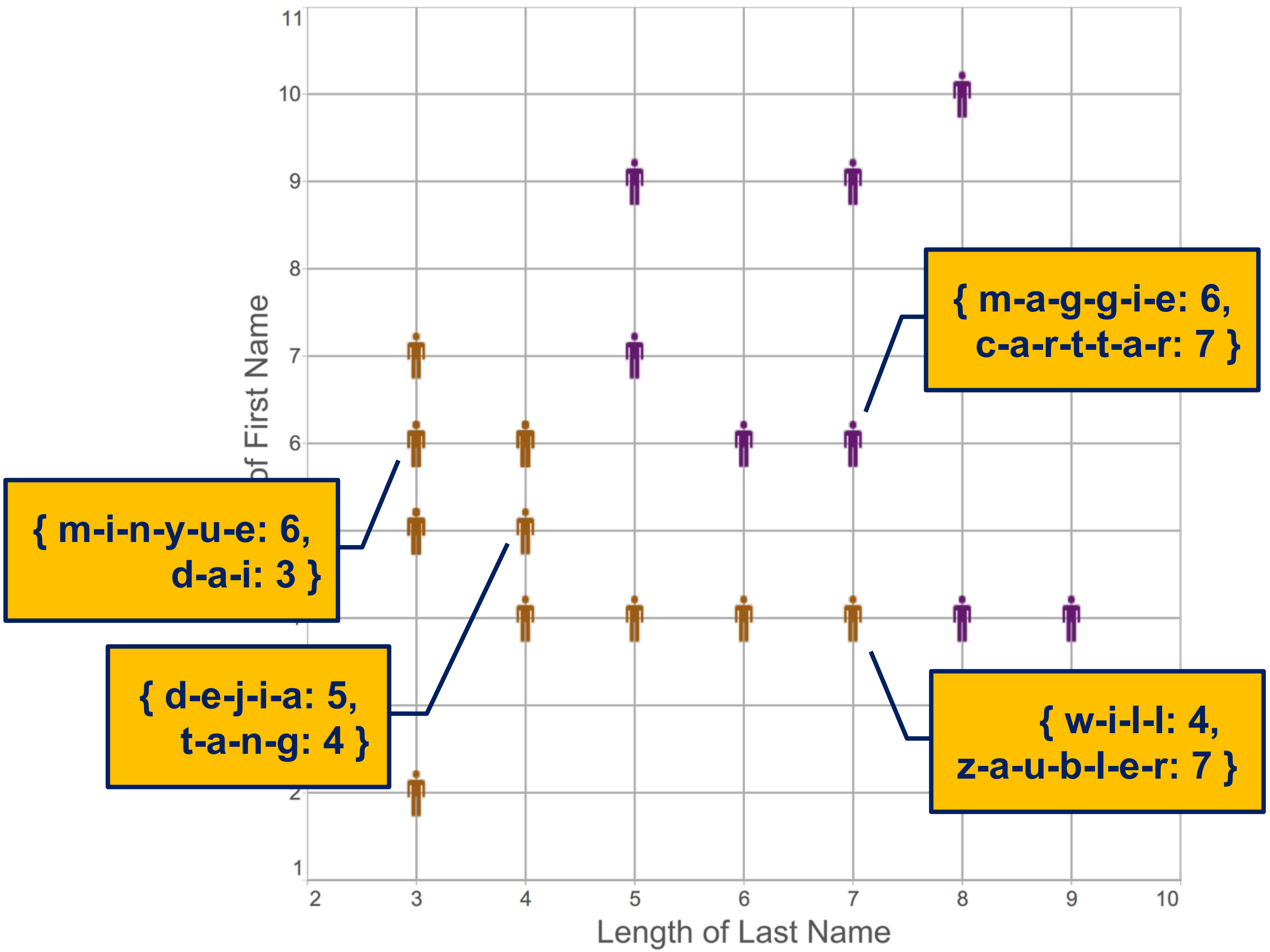
Announcements

- Thanks for your flexibility on Monday before break
- By popular request, today will be a split class:
  - Part 1: Introduction to SVMs
  - Part 2: Final Project Workshop

Outline

- Maximal margin classifier
- Support vector classification
  - 2 classes, linear boundaries
  - 2 classes, nonlinear boundaries
- Multiple classes
- Comparison to other methods
- Lab

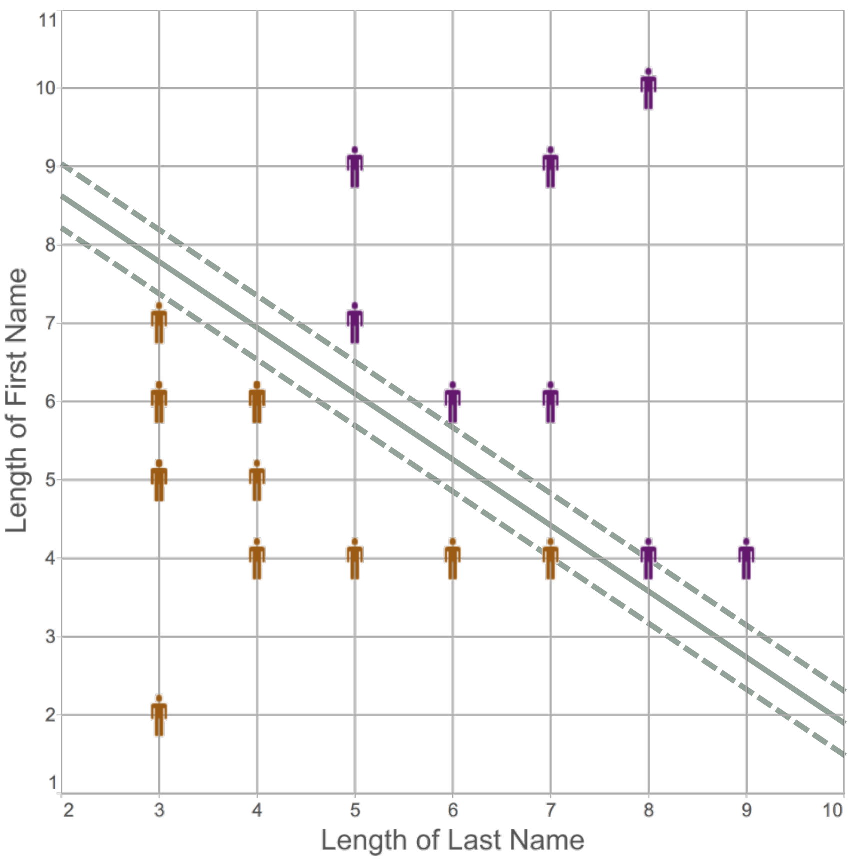
Toy example



{ m-a-g-g-i-e: 6,
c-a-r-t-t-a-r: 7 }
{ m-i-n-y-u-e: 6,
d-a-i: 3 }
{ d-e-j-i-a: 5,
t-a-n-g: 4 }
{ w-i-l-l: 4,
z-a-u-b-l-e-r: 7 }
of First Name
Length of Last Name
1
2
6
7
8
9
10
11
2
3
4
5
6
7
8
9
10

Maximal margin classifier

- **Claim**: if a separating hyperplane exists, there are infinitely many of them (why?)
- **Big idea**: pick the one that gives you the **widest margin** (why?)

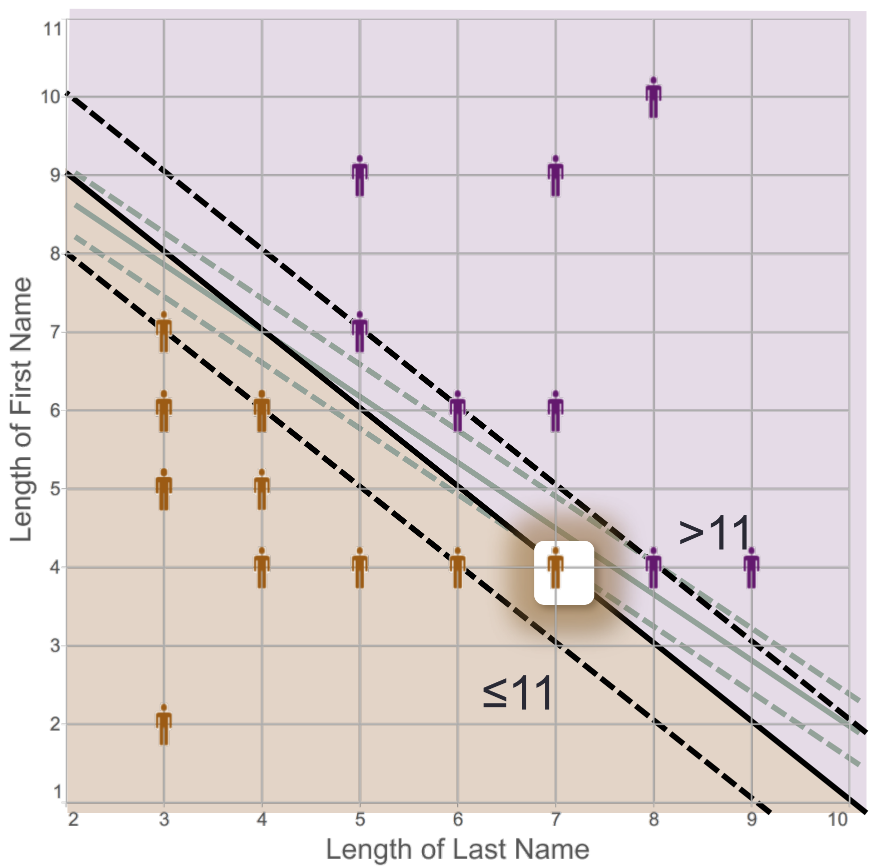


Bigger margin = **more confident**

Discussion

- **Question:** what's wrong with this approach?
- **Answer:** sometimes the **margin is tiny** (and therefore prone to overfitting), and other times the data **there is no hyperplane** that perfectly divides the data

Support vector classifier



- **Big idea**: might prefer to **sacrifice a few** if it enables us to perform better on the rest
- Can allow points to **cross the margin**, or even be completely **misclassified**

Support vector classifier (math)

- **Goal**: maximize the margin M
- Subject to the following constraints:

$$\underbrace{y_i\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)}_{\perp \text{ distance from the } i^{th} \text{ obs. to the hyperplane}} \geq M\boxed{\left(1-\varepsilon_i\right)}$$

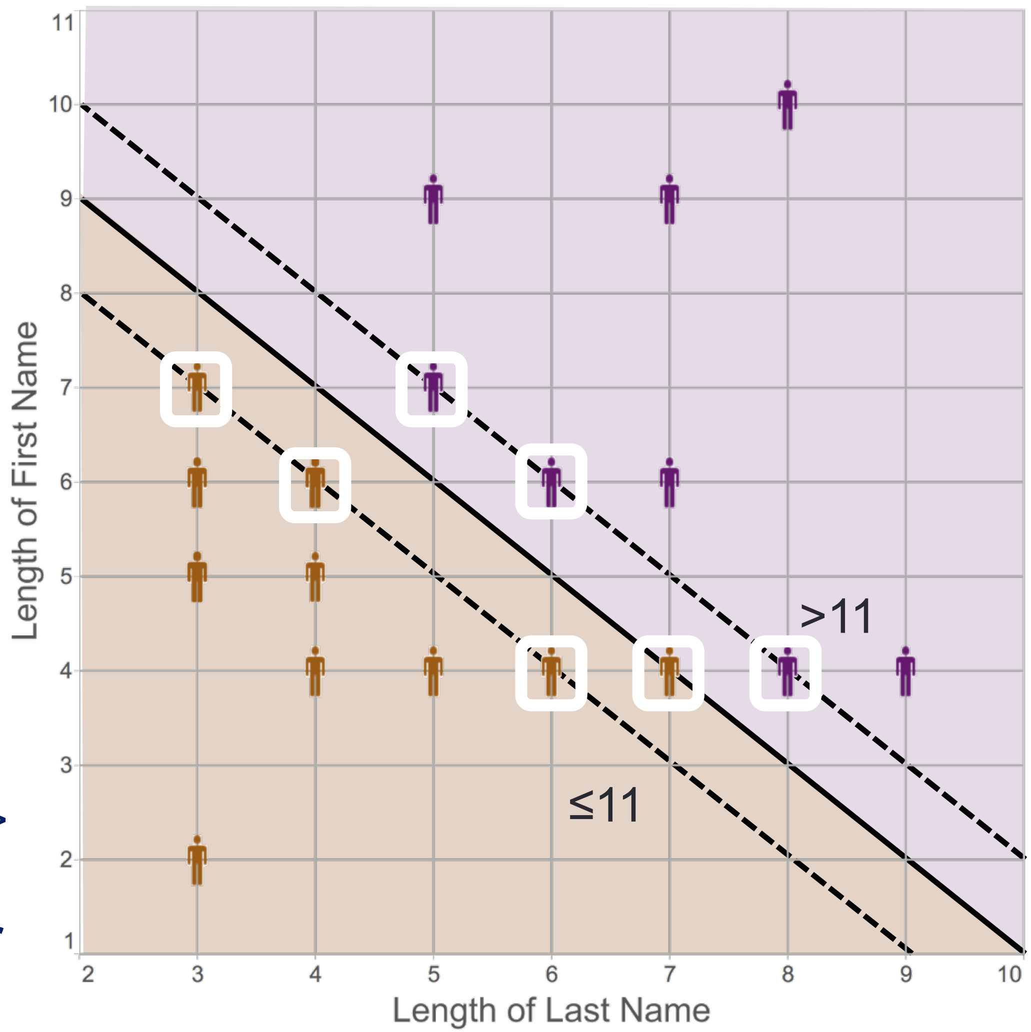
"slack variables"

$$\boxed{\varepsilon_i \geq 0,} \boxed{\sum_{i=1}^{n} \varepsilon_i \leq C}$$

no one gets rewarded for being extra far from the margin

We only have so much "slack" to allocate across all the observations

Support vectors



Decision rule is based only on the support vectors => SVC **is robust to strange behavior** far from the hyperplane!

A more general formulation

- **Goal**: maximize the margin M
- Subject to the following constraints:

$$\underbrace{y_i\left(\beta_0+\beta_1 x_{i1}+\ldots+\beta_p x_{ip}\right)}_{\perp \text{ distance from the } i^{th} \text{ obs. to the hyperplane}} \geq M\left(1-\varepsilon_i\right)$$

$$\varepsilon_i \geq 0, \quad \sum_{i=1}^{n} \varepsilon_i \leq C$$

- **Fun fact**: the solution to this maximization can be found using the **inner products** of the observations

Dot product = measure of similarity

- The **dot product** of 2 (same-length) vectors:

$$\begin{bmatrix} a_1 \\ \vdots \\ a_i \end{bmatrix} \bullet \begin{bmatrix} b_1 \\ \vdots \\ b_i \end{bmatrix} = \begin{bmatrix} a_1 \cdots a_i \end{bmatrix} \times \begin{bmatrix} b_1 \\ \vdots \\ b_i \end{bmatrix}$$

$$\text{Geometric} = \|A\|\|B\|\cos(\theta)$$

$$\text{Algebraic} = \sum_{j=1}^{i} a_j b_j$$

A more general formulation

- We can rewrite the linear support vector classifier as:

Only nonzero at support vectors

$$f(x) = \beta_0 + \sum_{i=1}^{n} \boxed{\alpha_i} (x \cdot x_i)$$

$$\rightarrow f(x) = \beta_0 + \sum_{i \in S} \alpha_i (x \cdot x_i)$$

- The **dot product** is just one way to measure the similarity
- In general, we call any such similarity measure a **kernel***

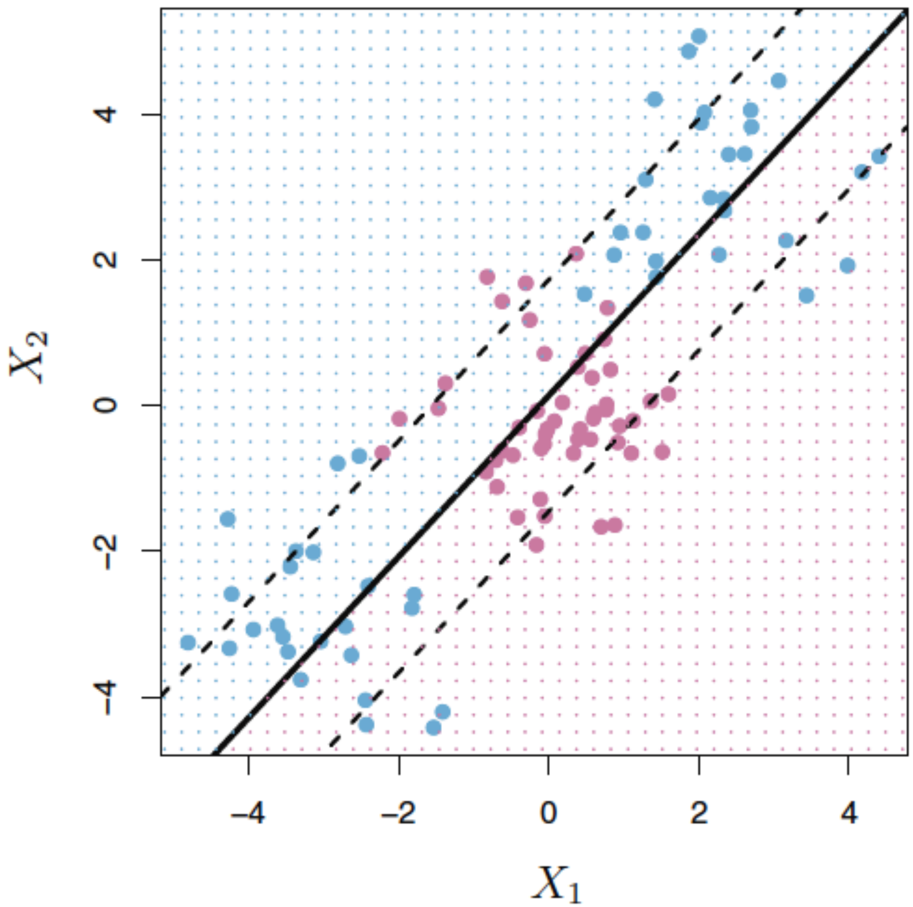
$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i)$$

*which is why SVMs and related measures are often referred to as "kernel methods"

Other kernels

We've seen a linear kernel (i.e. the classification boundary is **linear** in the features)

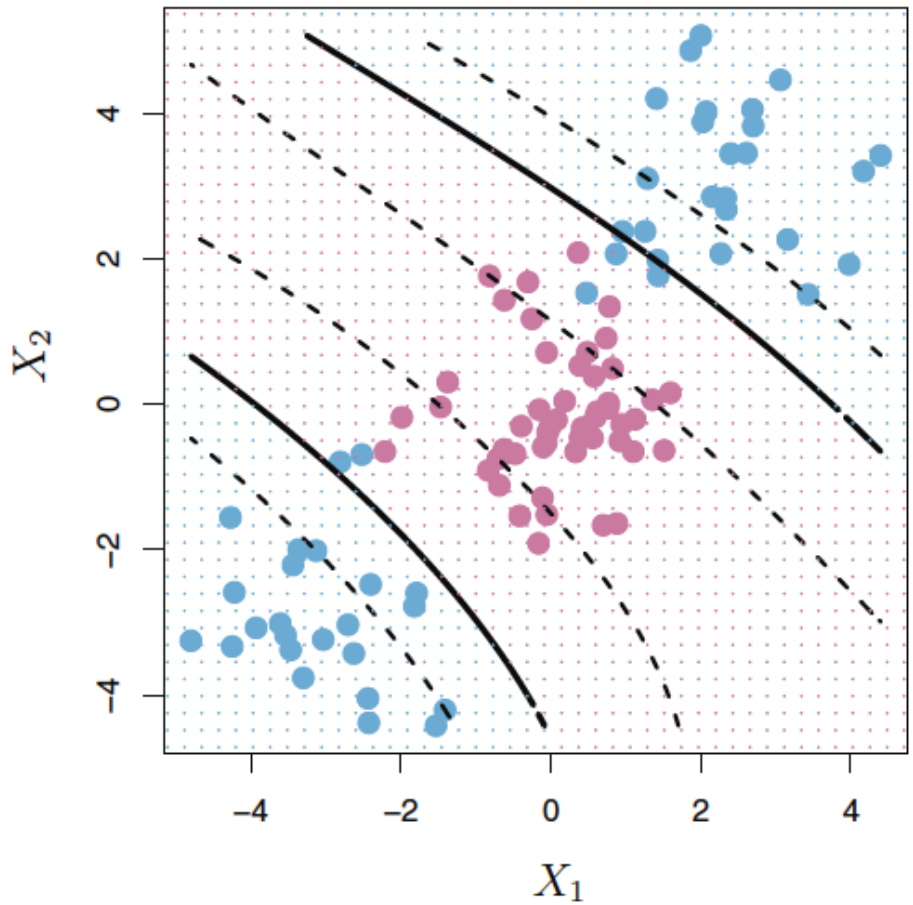
$$K\left(x_i, x_{i'}\right) = \sum_{j=1}^{p} x_{ij} x_{i'j}$$



Other kernels

We could also have a **polynomial** kernel

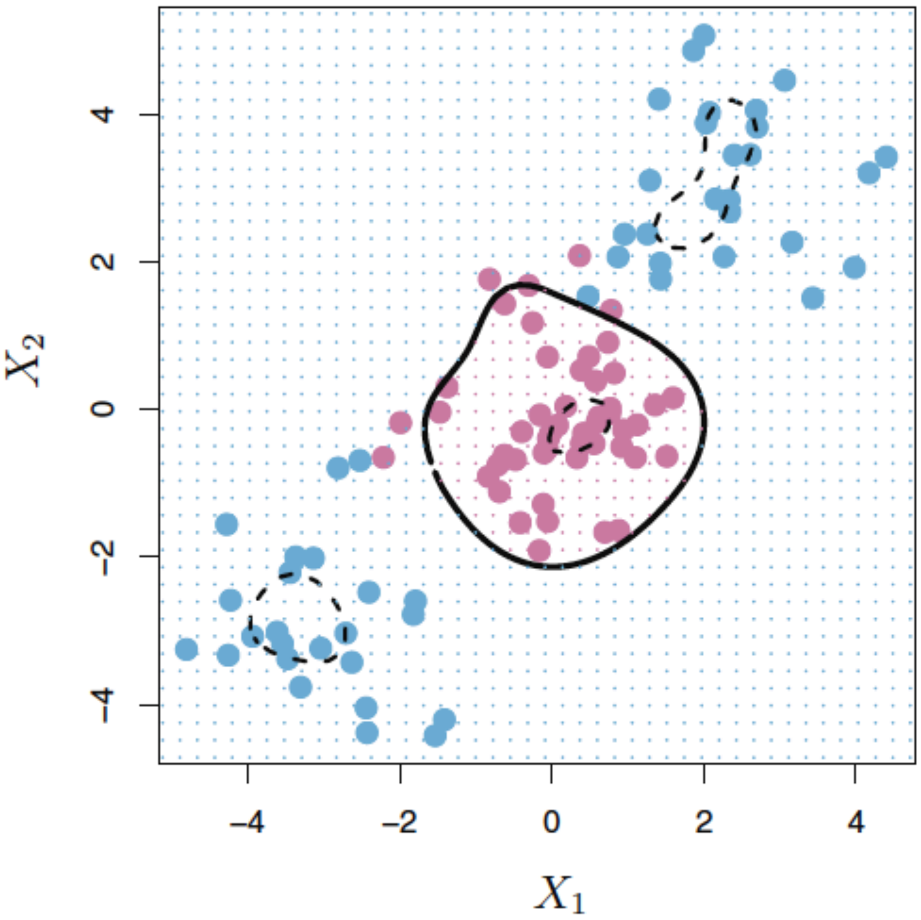
$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d$$



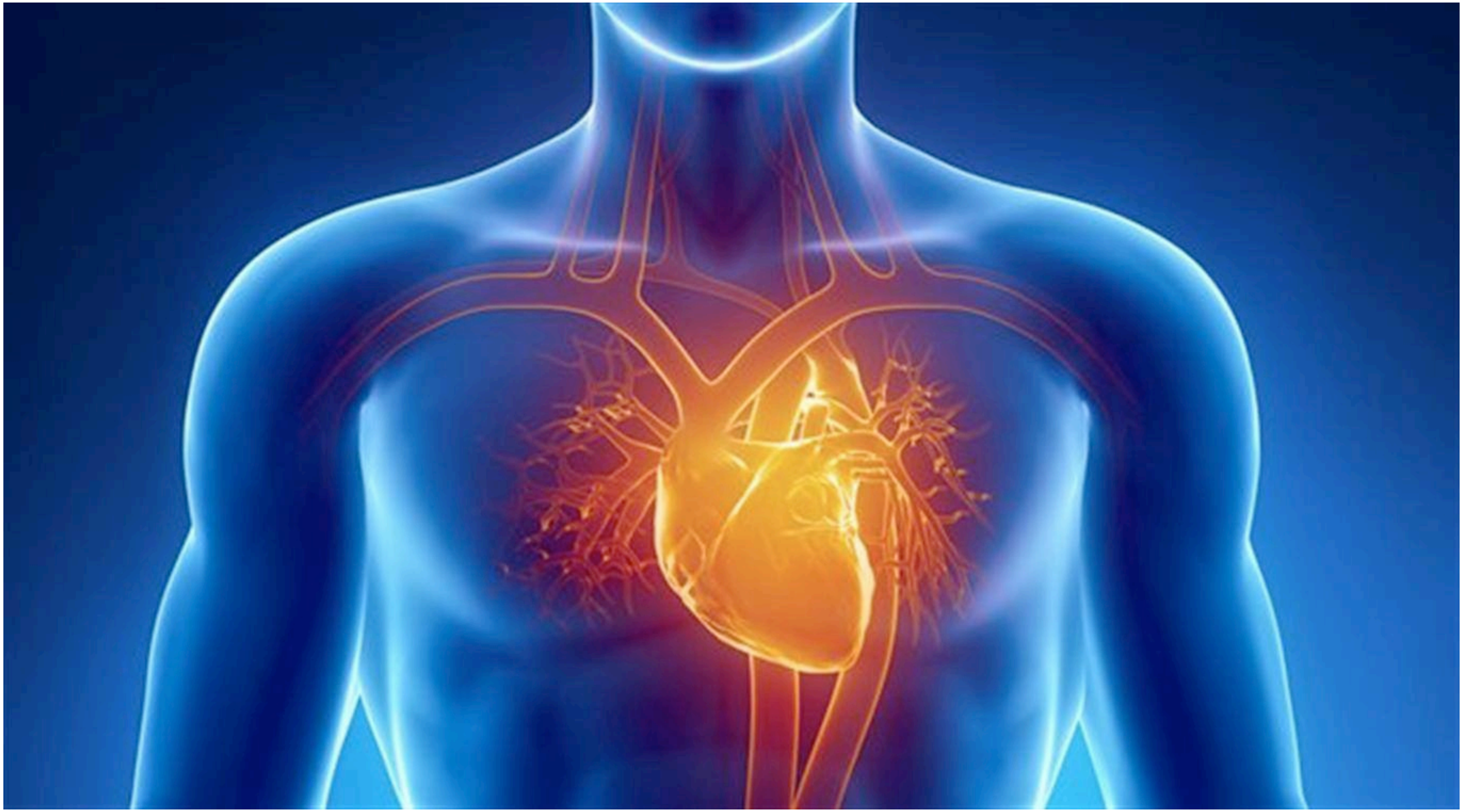
Other kernels

Or even a
radial kernel

$$K\left(x_i, x_{i'}\right) = e^{\left(-\gamma \sum_{j=1}^{p} \left(x_{ij} - x_{i'j}\right)^2\right)}$$



Flashback: `Heart` dataset



Application: `Heart` dataset

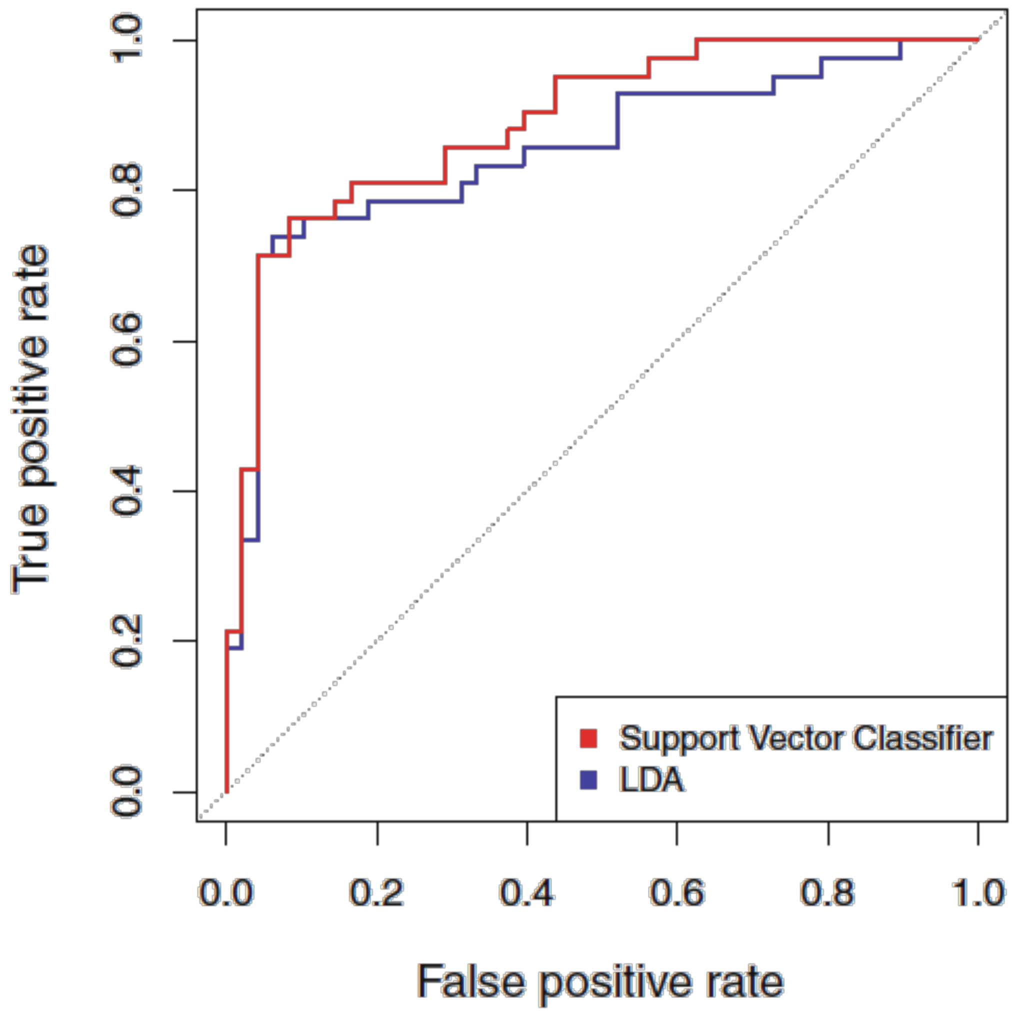
- **Goal:** predict whether an individual has heart disease on the basis of 13 variables: `Age`, `Sex`, `Chol`, etc.

- 297 subjects, randomly split into 207 training and 90 test observations

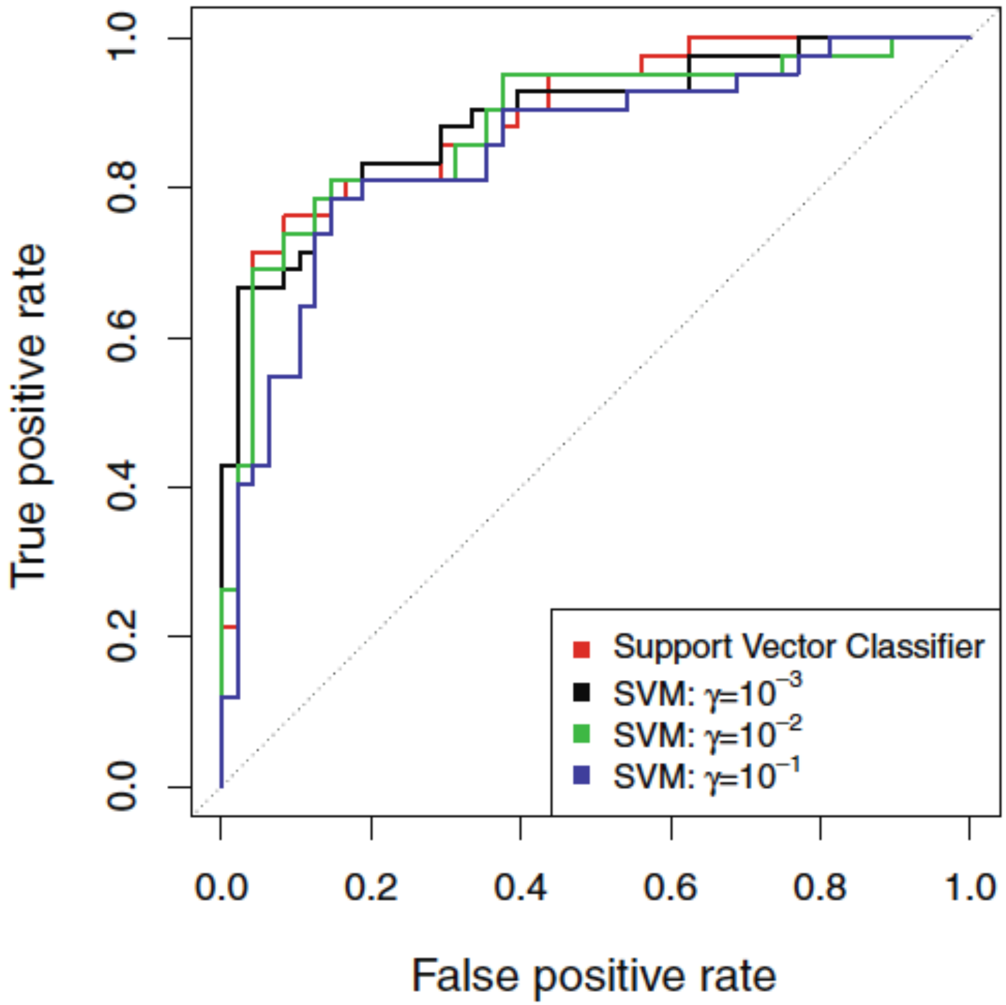
- **Problem:** no good way to plot in 13 dimensions

- **Solution:** ROC curve

ROC curve: LDA vs. SVM, linear kernel



ROC curve: SVM, linear vs. radial kernel



Coming up

- Wednesday**: Multiclass** support vector machines
- FP2 due tonight by 11:59pm
- FP3 out this afternoon