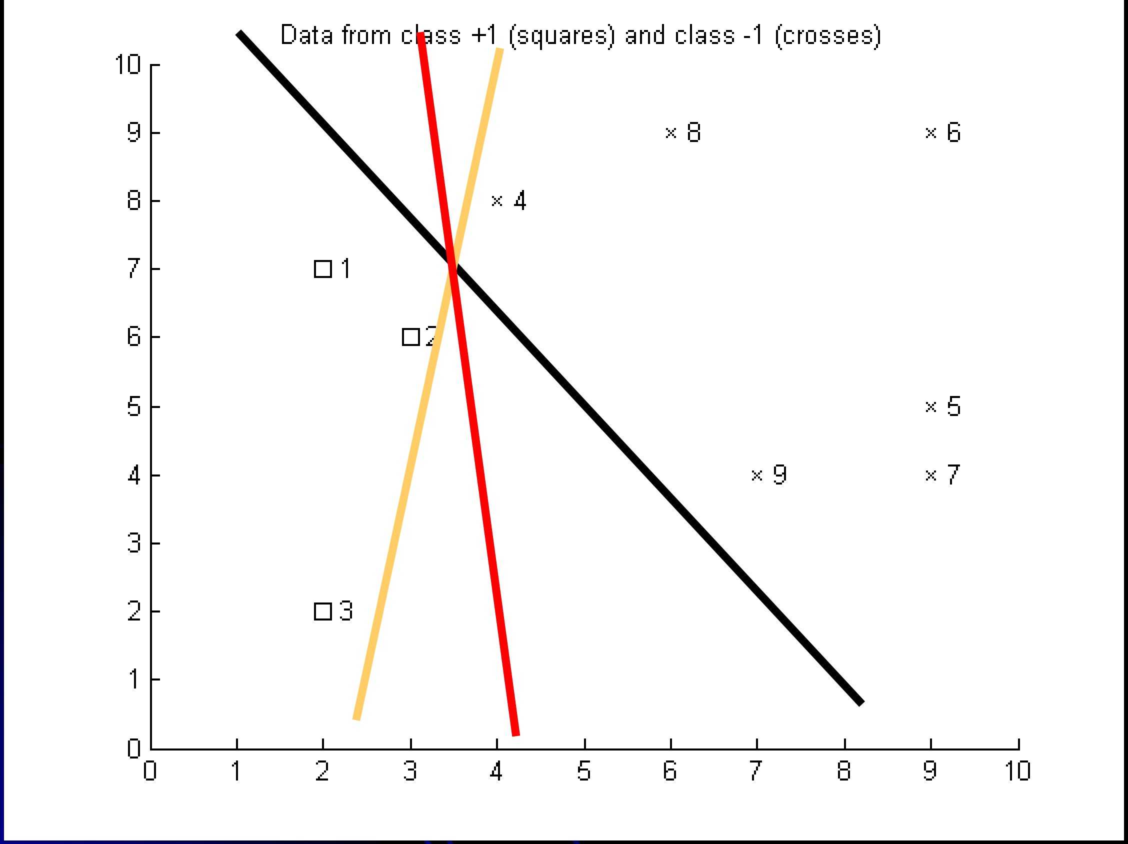
CSSE463: Image Recognition Day 14

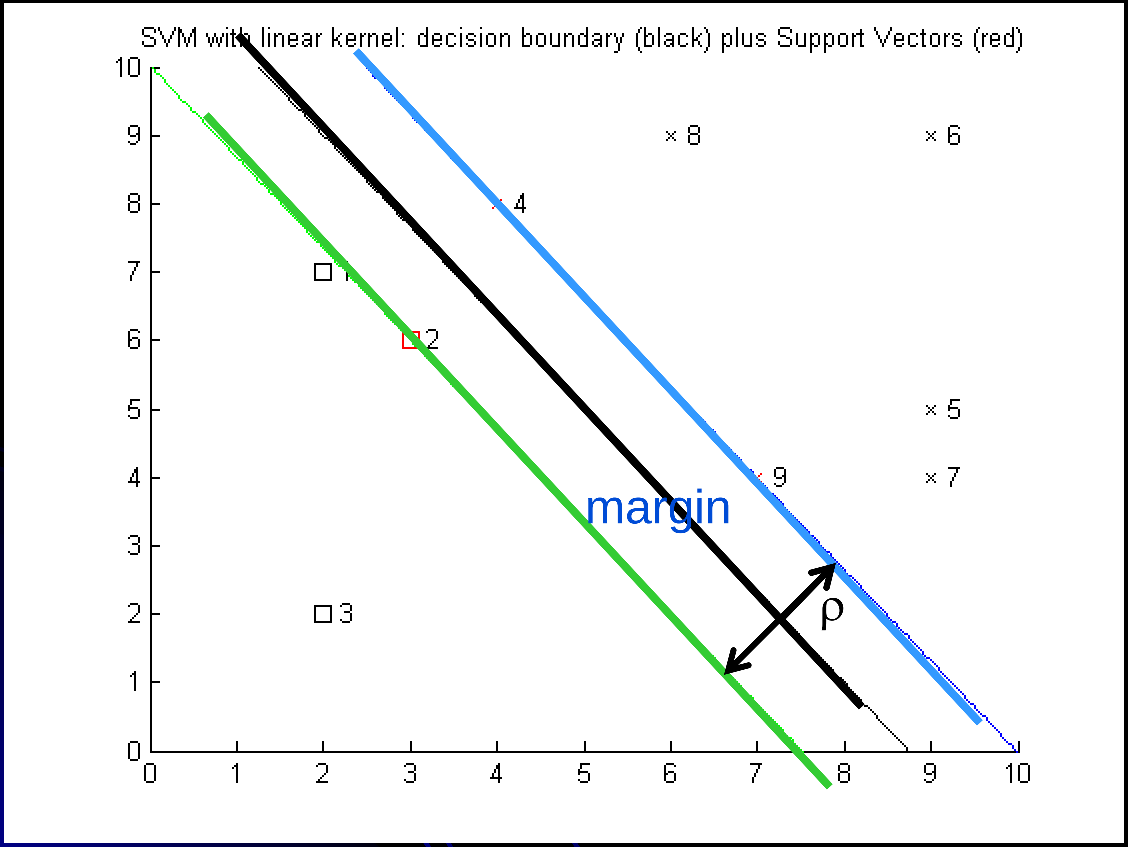
- Market analysis headline:
  - "The image recognition market is estimated to grow from $9.65 billion in 2014 to $25.65 billion by 2019."
- Lab due Weds.
  - These solutions assume that you don't threshold the shapes.ppt image:
    - Shape1: elongation = 1.632636, C1 = 19.2531, C2 = 5.0393
- Feedback on midterm plus/delta
    - *Projects/labs reinforce theory; interesting examples, topics, presentation; favorite class; enjoying*
    - Lecture and assignment pace OK or slightly off.
- This week:
  - Tuesday: Support Vector Machine (SVM) Introduction and derivation
  - Thursday: Project info, SVM demo
  - Friday: SVM lab

SVMs: “Best” decision boundary



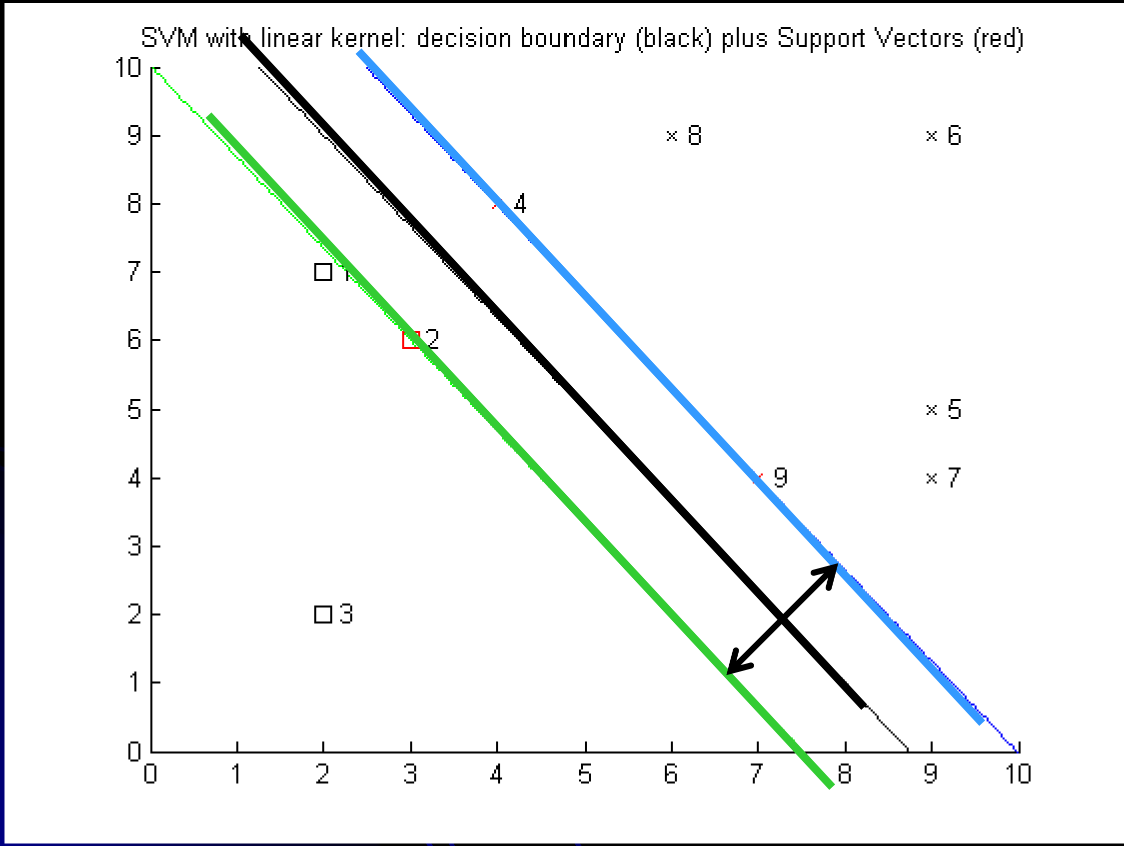
- Consider a 2-class problem
- Start by assuming each class is linearly separable
- There are many separating hyperplanes…
- Which would you choose?

SVMs: "Best" decision boundary



- The "best" hyperplane is the one that *maximizes the margin, ρ,* between the classes.
- Some training points will always lie on the margin
  - These are called *"support vectors"*
  - #2,4,9 to the left
- Why does this name make sense intuitively?

Support vectors



- The support vectors are the toughest to classify
- What would happen to the decision boundary if we moved one of them, say #4?
- A different margin would have maximal width!

Problem

- Maximize the margin width
- while classifying all the data points correctly…

Mathematical formulation of the hyperplane

- On paper
- Key ideas:
  - Optimum separating hyperplane:
  - Distance to margin:
  - Can show the margin width =
  - Want to maximize margin

Finding the optimal hyperplane

- We need to find w and b that satisfy the system of inequalities:

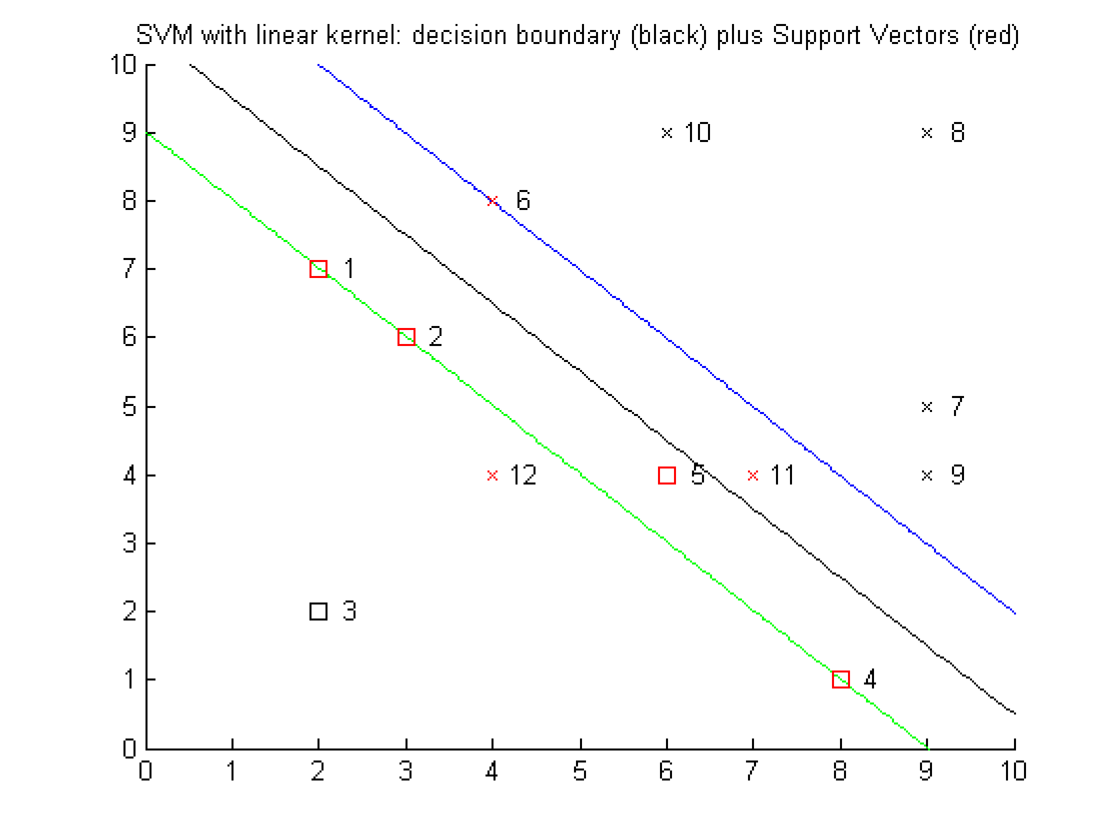
$$d_i(w^T x_i + b) \geq 1 \;\; for \; i = 1,2,\ldots.N$$

- where w minimizes the cost function:

$$\phi(w) = \frac{1}{2} w^T w$$

- (Recall that we want to minimize $||w_0||$, which is equivalent to minimizing $||w_{op}||^2 = w^T w$)
- Quadratic programming problem
  - Use Lagrange multipliers
  - Switch to the dual of the problem

Non-separable data



- Allow data points to be misclassifed
- But assign a cost to each misclassified point.
- The cost is bounded by the parameter C (which you can set)
- You can set different bounds for each class. Why?
  - Can weigh false positives and false negatives differently

Can we do better?

- Cover's Theorem from information theory says that we can map nonseparable data in the input space to a feature space where the data is separable, with high probability, if:
  - The mapping is nonlinear
  - The feature space has a higher dimension
- The mapping is called a *kernel function*.
  - Replace every instance of $x_i x_j$ in derivation with $K(x_i x_j)$
  - Lots of math would follow here to show it works
- Example:
  - separate x_1 XOR x_2 by adding a dimension $x_3 = x_1 x_2$

Most common kernel functions

- Polynomial
- Gaussian Radial-basis function (RBF)
- Two-layer perceptron

$$K(x, x_i) = (x^T x_i + 1)^p$$

$$K(x, x_i) = \exp\left(-\frac{1}{2\sigma^2}\left\|x - x_i\right\|^2\right)$$

$$K(x, x_i) = \tanh\left(\beta_0 x^T x_i + \beta_1\right)$$

- You choose p, σ, or β_i
- My experience with real data: **use Gaussian RBF!**

Easy — Difficulty of problem — Hard

p=1, p=2, — higher p — RBF

Q5