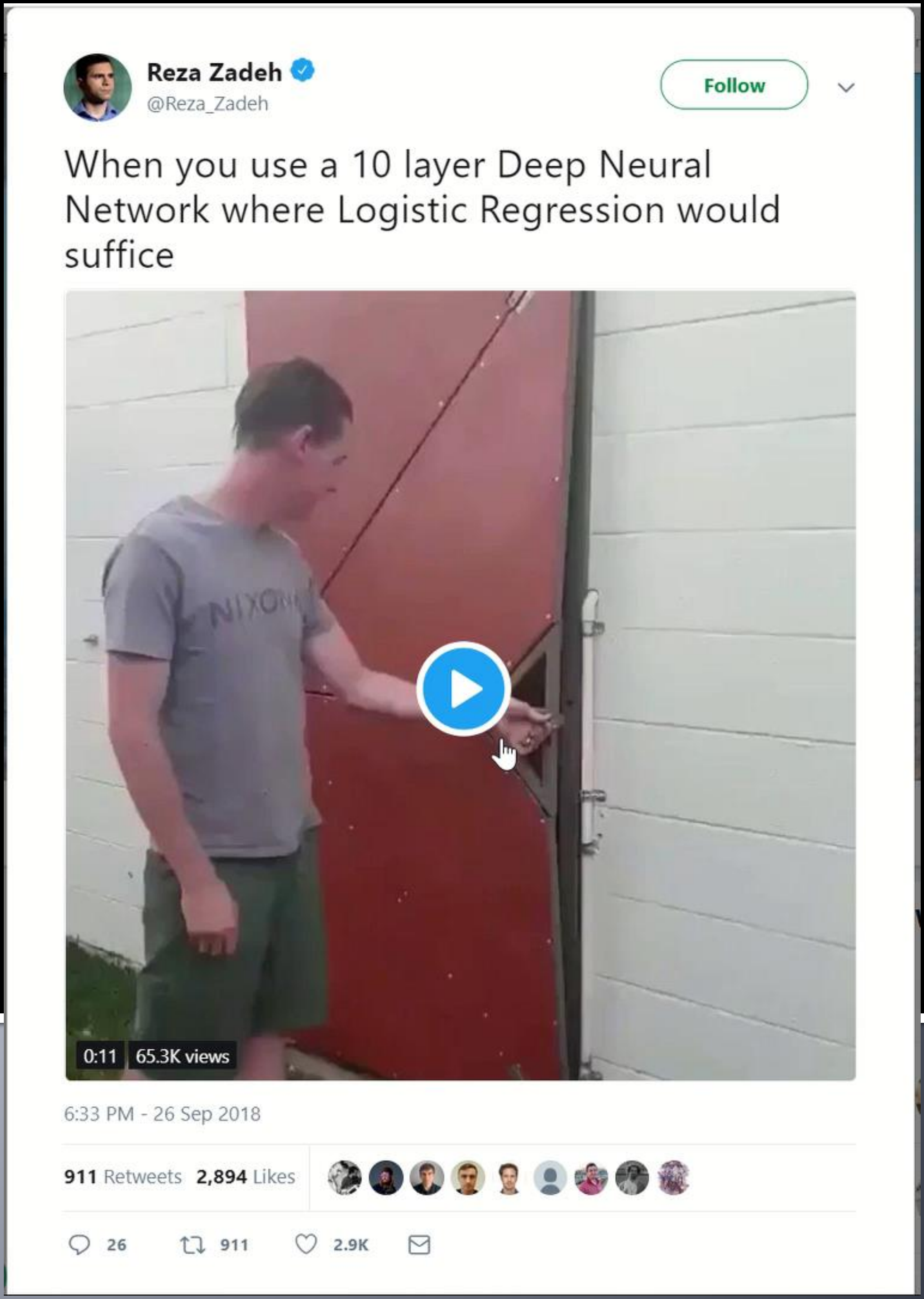
INFO 251: Applied Machine Learning

Logistic Regression



Announcements

- Assignment 4 will be posted today
- Quiz 1 scheduled for next Tuesday, 9:40-10:20
  - 10-15 multiple choice and short-answer questions
  - Contact course staff via Piazza if you can't make this time, we are arranging a "make-up" time at 9:00-9:40 on the same day.

Key Concepts (last lecture)

- Overfitting
- Regularization: Intuition
- Regularization: Cost function adjustment
- Ridge
- Lasso
- Cross-validation of regularization hyperparameters
- Coefficient plots

Gradient descent: Example Quiz Question

- To ensure that your gradient descent algorithm is properly converging to a minimum:
  1. Plot $J(\theta)$ as a function of θ, and ensure $J(\theta)$ is decreasing
  2. Plot $J(\theta)$ as a function of number of iterations, and ensure $J(\theta)$ is decreasing
  3. Plot $J(\theta)$ as a function of θ, and make sure $J(\theta)$ is convex
  4. Plot $J(\theta)$ as a function of learning rate R, and make sure $J(\theta)$ in monotonic (either constantly increasing or constantly decreasing) in R

Course Outline

- Causal Inference and Research Design
  - Experimental methods
  - Non-experiment methods
- Machine Learning
  - Design of Machine Learning Experiments
  - **Linear Models and Gradient Descent**
  - Non-linear models
  - Neural models
  - Unsupervised Learning
  - Practicalities, Fairness, Bias
- Special topics

Key Concepts (this lecture)

- Logistic regression
- Simplified sigmoid cost function
- Odds ratios
- Overfitting revisited
- Support vector machines
- Hard vs. soft margins
- Kernel functions

Outline

- Logistic regression (inference)
- Logistic regression (prediction and gradient descent)
- Support vector machines
- Kernels

Logistic regression: Basics

- Logistic regression
  - Models the (linear) relationship between one or more independent variables and one binary dependent variable
  - As with linear regression, can be used for inference and prediction; used to predict (and classify) binary outcomes

Inference	**Prediction**
What is the effect of an additional year of schooling on whether an individual is eligible for welfare?	Do we predict that an individual with 6 years of education will be eligible for welfare?
What caused the server to go down last week?	Will the server go down this week?
How big a factor is “home court advantage” in whether our team will win or lose?	Are we going to win this week?

Logistic Regression: Idea

- Logistic Regression: Model
  - The logistic regression model assumes that the independent variables have a linear relationship with the logit transformation of the dependent variable
    - i.e., $\text{logit}(y) = \alpha + \beta X + \cdots$

Logistic Regression: Idea

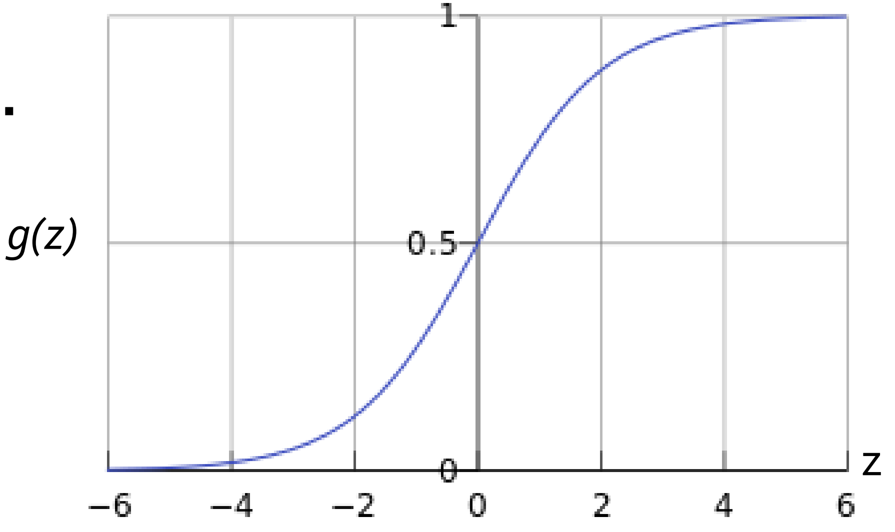
- Logit transformation maps probabilities to log of odds ratios
  - Odds ratio: probability success / probability failure, or $\frac{p}{1-p}$
  - Example: Probability success = 0.8
    - Odds ratio is 4
    - "Odds of success are 4 to 1"

- In other words:
  - $\mathrm{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta X + \cdots$
  - $p = \frac{e^{\alpha+\beta X+\cdots}}{1+e^{\alpha+\beta X+\cdots}}$

Logistic Regression: The logistic function

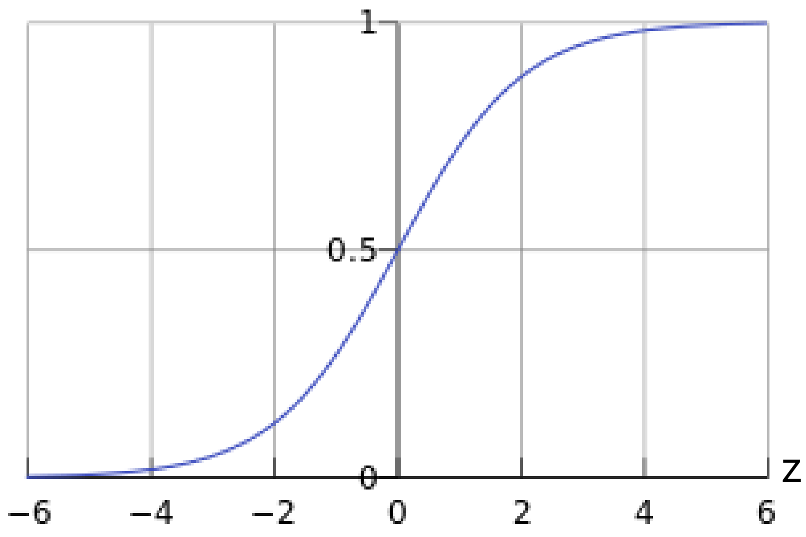
- Logistic (sigmoid) function: $g(z) = \frac{e^{z}}{e^{z}+1} = \frac{1}{1+e^{-z}}$
  - Transforms [-∞,+∞]=>[0,1]
  - Constrains output of our model between 0 and 1

- In logistic regression, $z = \alpha + \beta X + \ldots$
  - Logistic is thus: $g(z) = \frac{1}{1+e^{-(\alpha+\beta X+\cdots)}}$



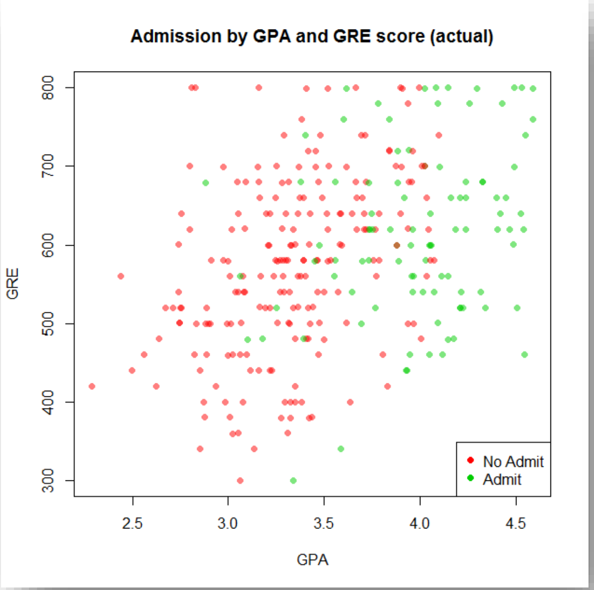
Logistic Regression: Decision Boundary

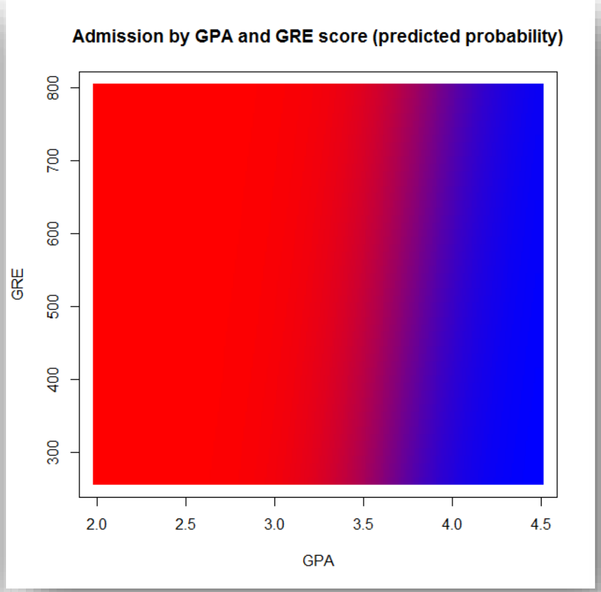
- How to interpret g(z)?
  - Probability that y=1
  - $P(y = 1|x: \alpha, \beta)$
- Simple classifier
  - Predict y=1 if g(z)≥0.5
  - Predict y=0 if g(z)<0.5
- How does this relate to values of z?
  - Predict y=1 if z≥0
  - Predict y=0 if z<0
  - Typically, $z = \alpha + \beta X + \cdots$

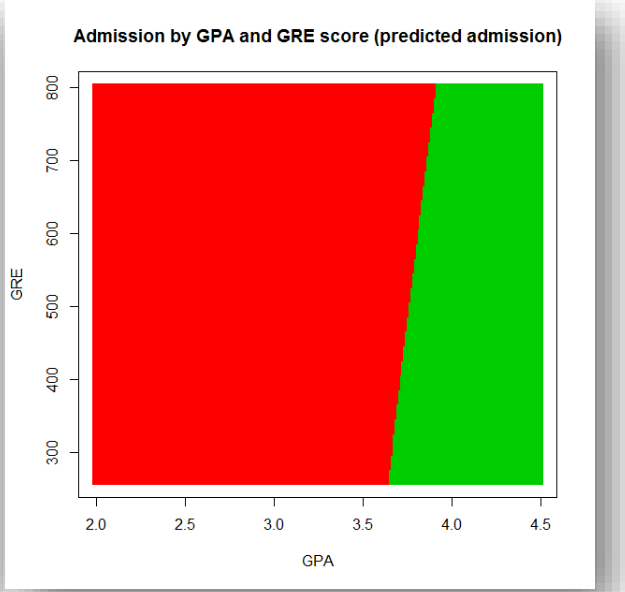


Logistic Regression: Example

- Example: admission vs. GRE and GPA
  1. Start with raw data
  2. Fit logistic regression
  3. Threshold converts g(z) to classification







Logistic Regression: Coefficients

- How do we interpret the coefficients from a logistic regression?
  - The coefficient tells you what change to expect in the *log odds ratio* of your dependent variable, for a one-unit increase in your independent variable.

- Ways to make this more intelligible
  - Convert from log odds ratio to odds ratio
    - $\exp(\beta)$
  - Convert from odds ratio to probability
    - $\frac{odds}{1+odds}$

Logistic Regression: Coefficients

- Example with no predictor variables
  - Likelihood of being honor student
    - $\text{logit}(honor_i) = \alpha + \epsilon_i$

```
Logistic regression                             Number of obs   =        200
                                                LR chi2(0)      =       0.00
                                                Prob > chi2     =          .
Log likelihood = -111.35502                     Pseudo R2       =     0.0000

------------------------------------------------------------------------------
         hon |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   intercept |   -1.12546   .1644101    -6.85   0.000    -1.447697   -.8032217
------------------------------------------------------------------------------
```

    - i.e., log(p/(1-p)) = -1.12546

  - Note that p = exp(-1.12546)/(1+exp(-1.12546)) = .245

```
        hon |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        151       75.50       75.50
          1 |         49       24.50      100.00
------------+-----------------------------------
      Total |        200      100.00
```

Logistic Regression: Coefficients

- Example with single predictor variable
  - Likelihood of honor student, by major
    - $\text{logit}(honor_i) = \alpha + \beta STEM_i + \epsilon_i$

```
Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       3.10
                                                  Prob > chi2     =     0.0781
Log likelihood = -109.80312                       Pseudo R2       =     0.0139
```

hon	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
stem	.5927822	.3414294	1.74	0.083	-.0764072	1.261972
intercept	-1.470852	.2689555	-5.47	0.000	-1.997995	-.9437087

  - exp(0.593) = 1.809
    - (this is the odds ratio)
    - (corresponds to p=0.644)
  - The odds radio can also be seen in the cross-tabs:
    - Odds for non-STEM: 0.23 (17/74)
    - Odds for STEM: 0.42 (32/77)
    - Odds for STEM 81% higher
      - 0.42 / 0.23 = 1.809
      - 0.644 / (1-0.644) = 1.809

	stem		
hon	**no**	**yes**	Total
0	74	77	151
1	17	32	49
Total	91	109	200

Outline

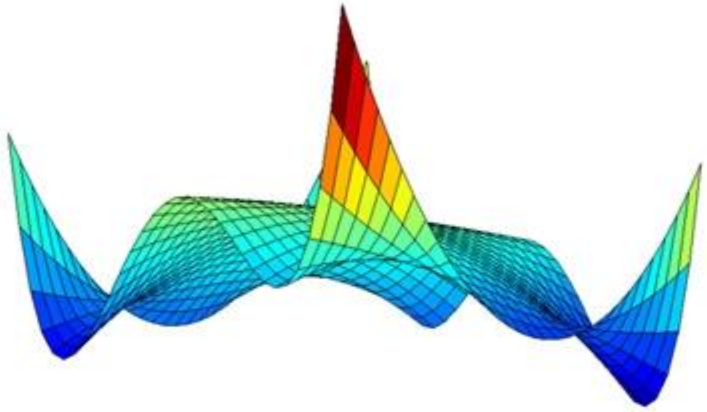
- Logistic regression (inference)
- **Logistic regression (prediction & gradient descent)**
- Support vector machines
- Kernels

Logistic Regression: General formulation

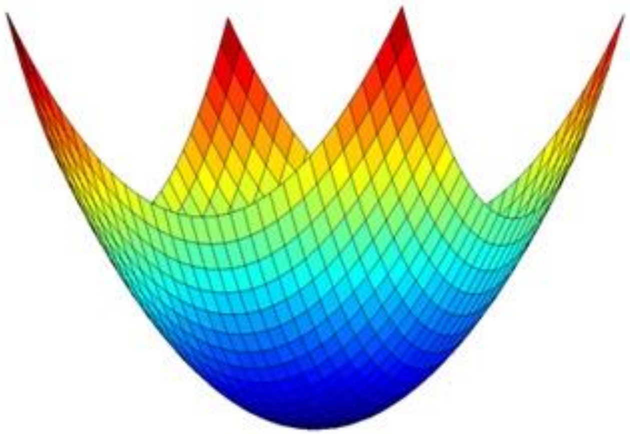
- Model ("hypothesis")
  - $P(Y_i = 1|x\colon \theta) = g(z) = \frac{1}{1+e^{-z}}$
- Parameters
  - θ are the parameters, often α, β
  - If $\theta = (\alpha, \beta)$, $\quad P(Y_i = 1) = \frac{1}{1+e^{-(\alpha+\beta X_i)}}$
- Cost Function
  - $J(\theta) = \frac{1}{N}\sum_{i=1}^{N} \text{Cost}(\hat{Y}_i, Y_i)$
  - (more on this shortly)
- Objective
  - $\min_{\theta} J(\theta)$

Logistic Regression: Cost function

- Cost Function
  - Linear regression: $J(\alpha, \beta) = \frac{1}{2N}\sum_{i=1}^{N}(Y_i - \alpha - \beta X_i)^2$
  - Why not $J(\alpha, \beta) = \frac{1}{2N}\sum_{i=1}^{N}\left(Y_i - \frac{1}{1+e^{-\alpha-\beta X_i}}\right)^2$

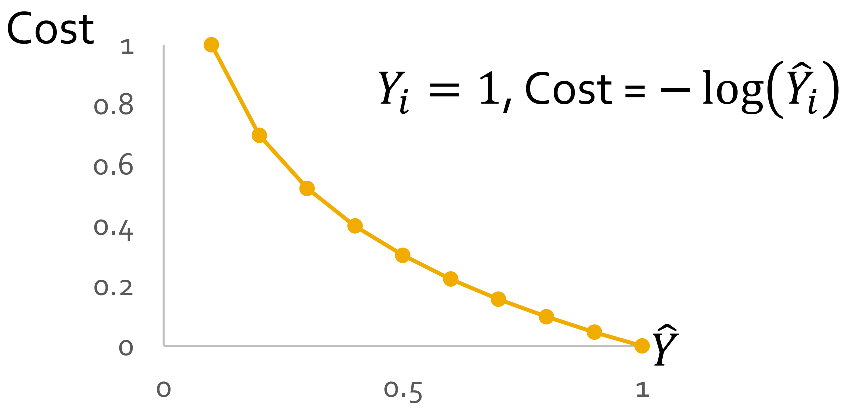


- Not convex ☹
  - Sigmoid function is complex, $J(\alpha, \beta)$ is not convex…
  - Susceptible to local minima, want to convert to something convex

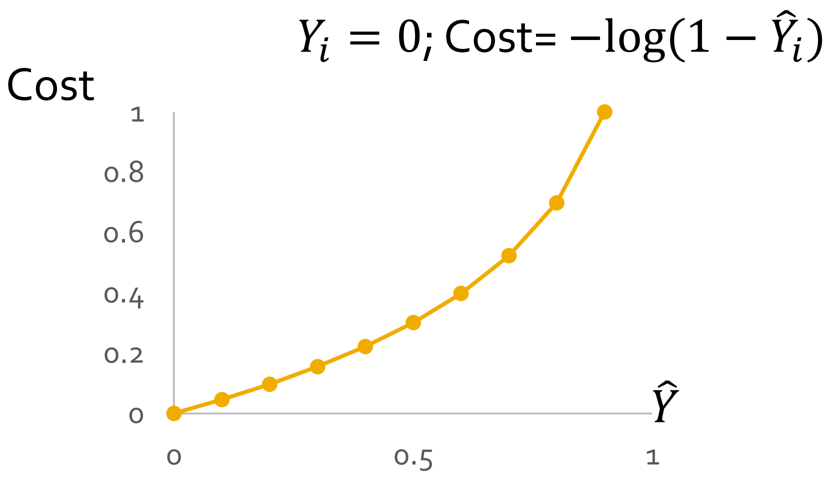


Logistic Regression: Cost function

- Cost Function (think of $\hat{Y}_i = \frac{1}{1+e^{-(\alpha+\beta X_i)}}$)
  - $\text{Cost}(\hat{Y}_i, Y_i) = \begin{cases} -\log(\hat{Y}_i) & \text{if } Y_i = 1 \\ -\log(1-\hat{Y}_i) & \text{if } Y_i = 0 \end{cases}$
  - $\text{Cost}(\hat{Y}_i, Y_i) = -Y_i \cdot \log(\hat{Y}_i) - (1-Y_i) \cdot \log(1-\hat{Y}_i)$



- This is convex:
  - If $Y_i = 1$, what is cost if $\hat{Y}_i = 1$? What if $\hat{Y}_i = 0$?
    - No cost if model predicts 1
    - Penalizes mistakes
  - If $Y_i = 0$, what is cost if $\hat{Y}_i = 1$? if $\hat{Y}_i = 0$?
    - No cost if model predicts 0
    - Penalizes mistakes



Logistic Regression: Gradient Descent

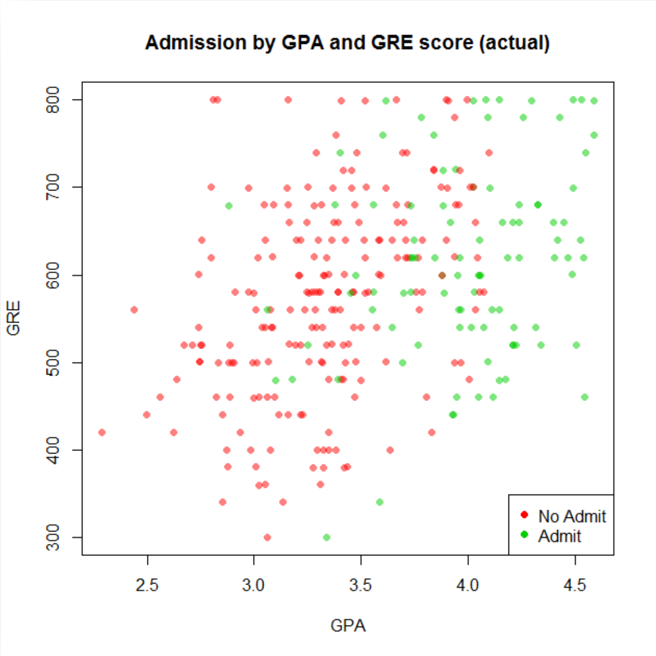
- Given the cost function $J(\theta)$, we now want to minimize:
  - $J(\theta) = -\frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \log \hat{Y}_i + (1 - Y_i) \log\left(1 - \hat{Y}_i\right)$
- Gradient Descent!
  - $\theta \leftarrow \theta - R\frac{\partial}{\partial\theta}J(\theta)$
- With revised cost function, $\frac{\partial}{\partial\theta}J(\theta) = -\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)X_i$
  - Note similarities to linear regression! But not identical:
  - Logistic regression: $\hat{Y}_i = \frac{1}{1+e^{-(\alpha+\beta X_i)}}$
- Gradient Descent Algorithm (logistic regression)
  - Repeat until convergence:
  - $\beta \leftarrow \beta + R\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)X_i$
  - in other words: $\beta \leftarrow \beta + R\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \frac{1}{1+e^{-(\alpha+\beta X_i)}}\right) \quad X_i$

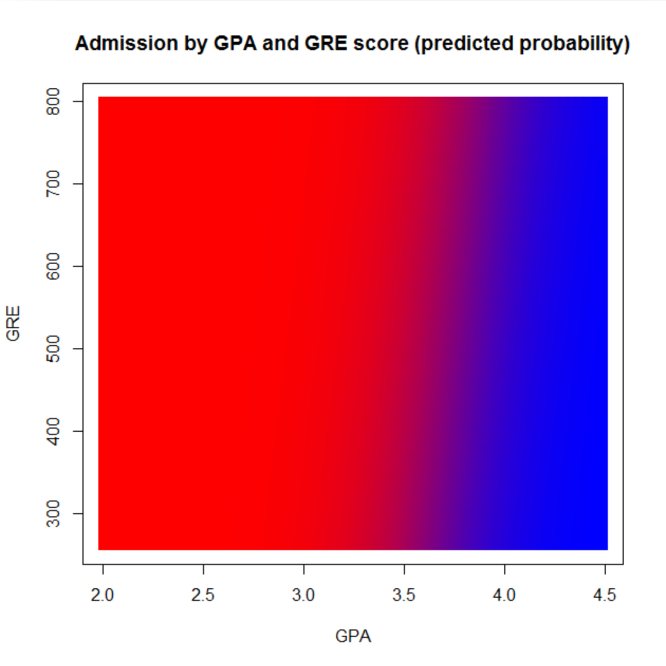
Outline

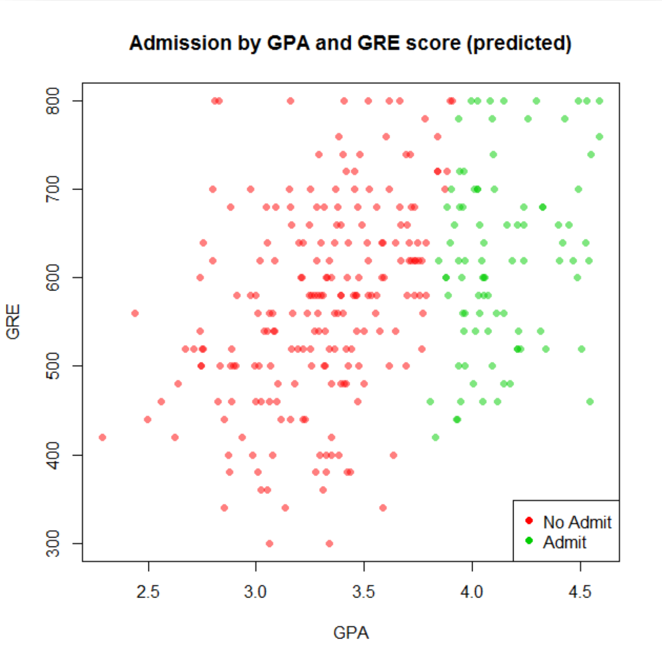
- Logistic regression (inference)
- Logistic regression (prediction and gradient descent)
- **Support vector machines**
- Kernels

Logistic Regression: Recap

- Compare actual vs. predicted values from our logistic regression

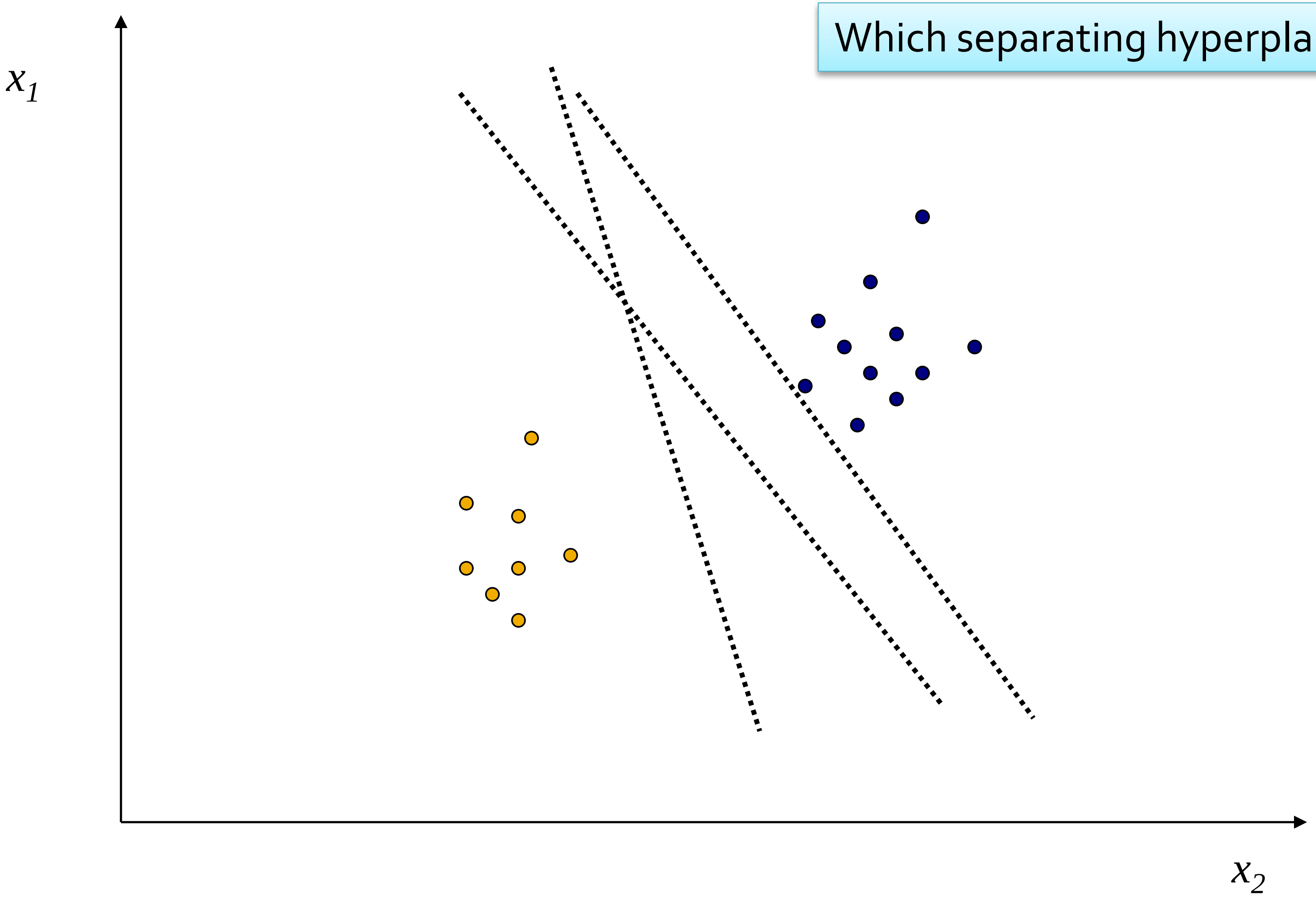




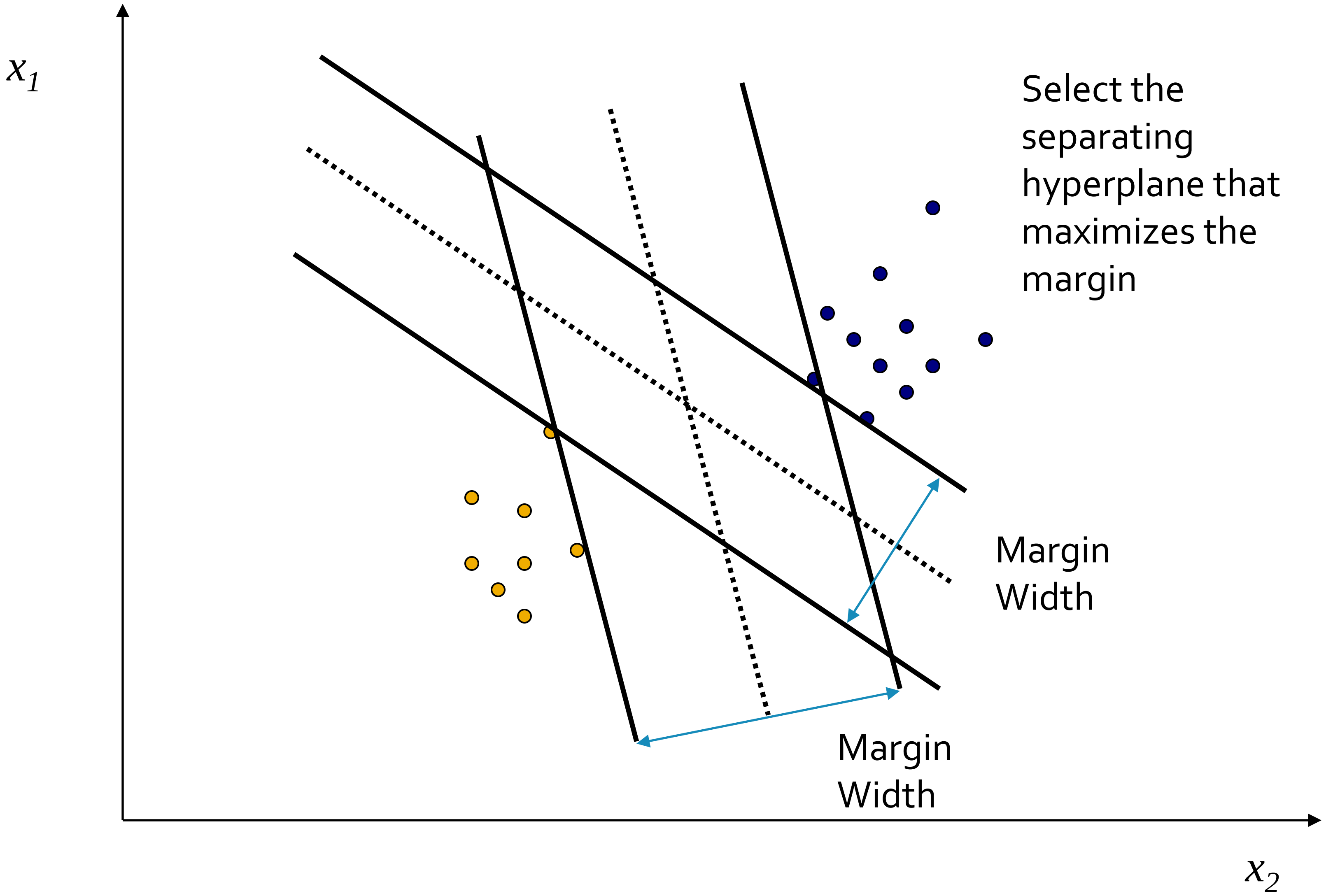


Support Vector Machines (SVM): Intuition

Which separating hyperplane?



SVM's objective: Maximize the Margin



Select the separating hyperplane that maximizes the margin

SVM Definition

- SVM defined by a separating plane
  - Represented by a weight vector *w*, and an intercept *b*
- Classifier function: $f(x) = sign(w^T x + b)$
- We can find an SVM classifier by solving the system of constraints (a quadratic programming problem):
  - $\max_{w,b}(\alpha)$ maximize the margin
  - where $w^T x - b \geq \alpha$ for points *x* in the first class
  - and $w^T x - b \leq -\alpha$ for points *x* in the second class
  - with $w^T w = 1$

- See Daume chapter 7

Soft-Margin SVM

- What if there is no separating hyperplane?
  - Introduce penalties ξ_i to mis-classifications
  - Helps prevent overfitting

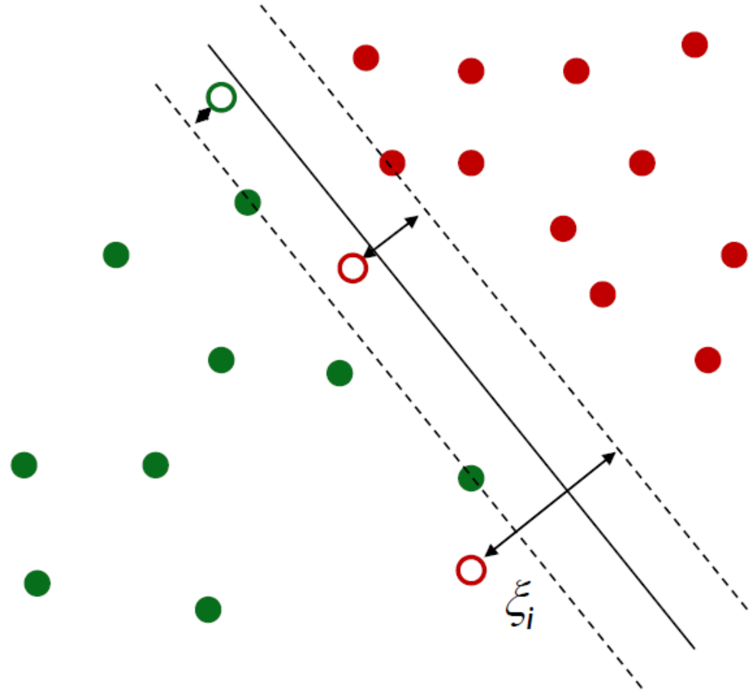


Image: John Canny

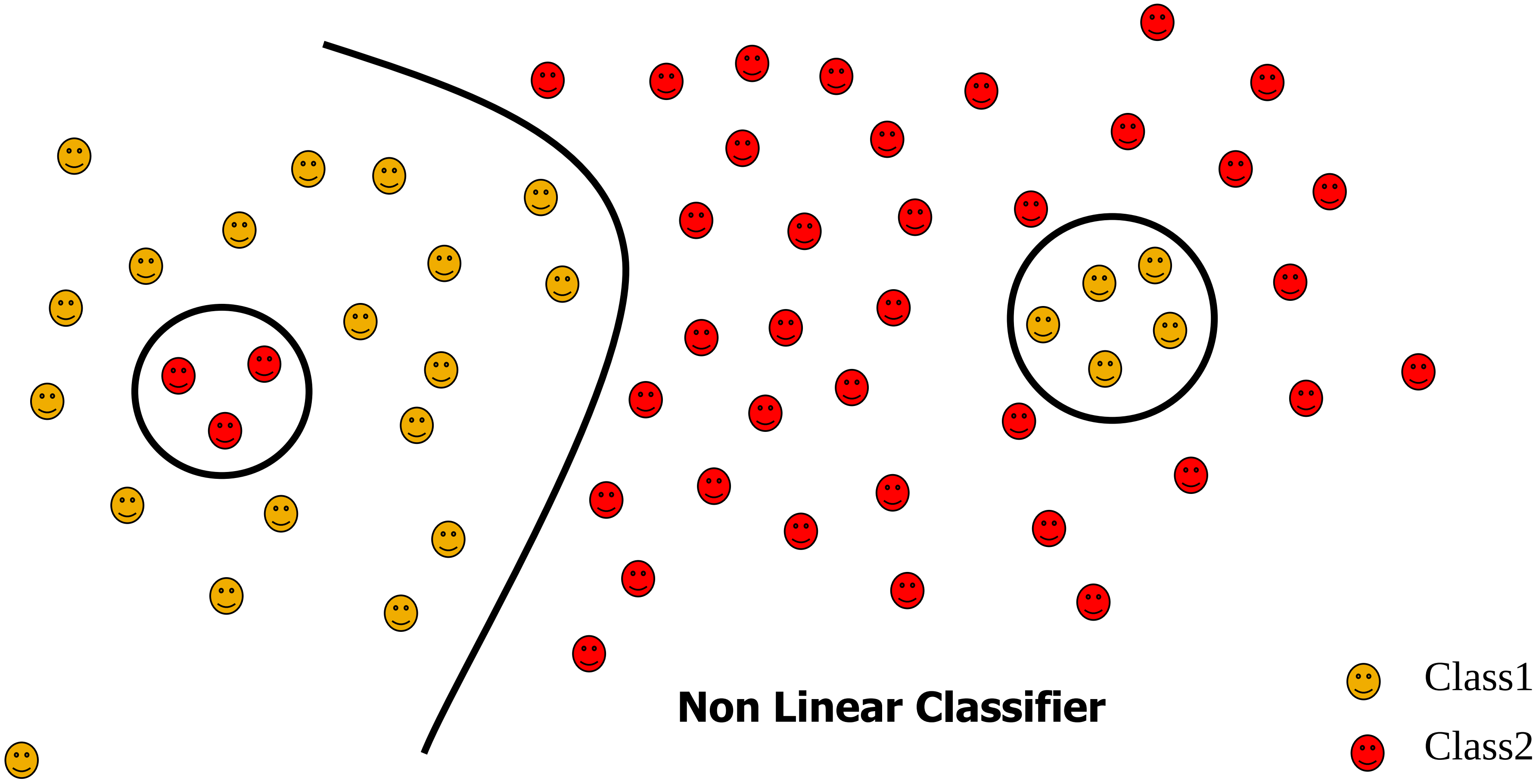
Linear models: Recap

- Linear models rely on some notion of a linear boundary (i.e., a hyperplane)
- But real-world data are typically not linearly separable
- Some classifiers just make a decision as to which class an object is in; others estimate class probabilities

Outline

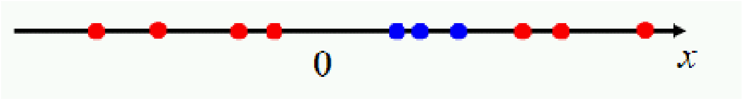
- Logistic regression (inference)
- Logistic regression (prediction and gradient descent)
- Support vector machines
- **Kernels**

Nonlinearly separable data

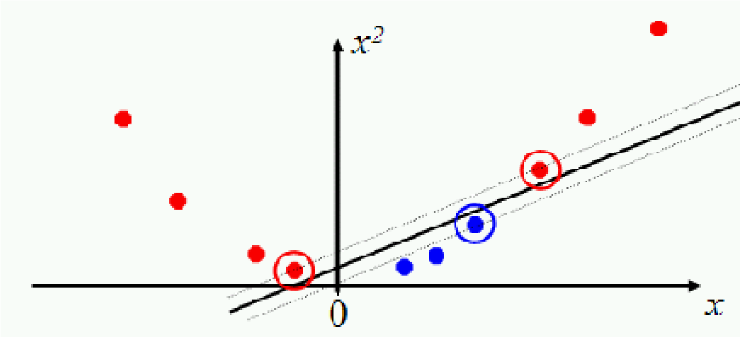


Extending linear models

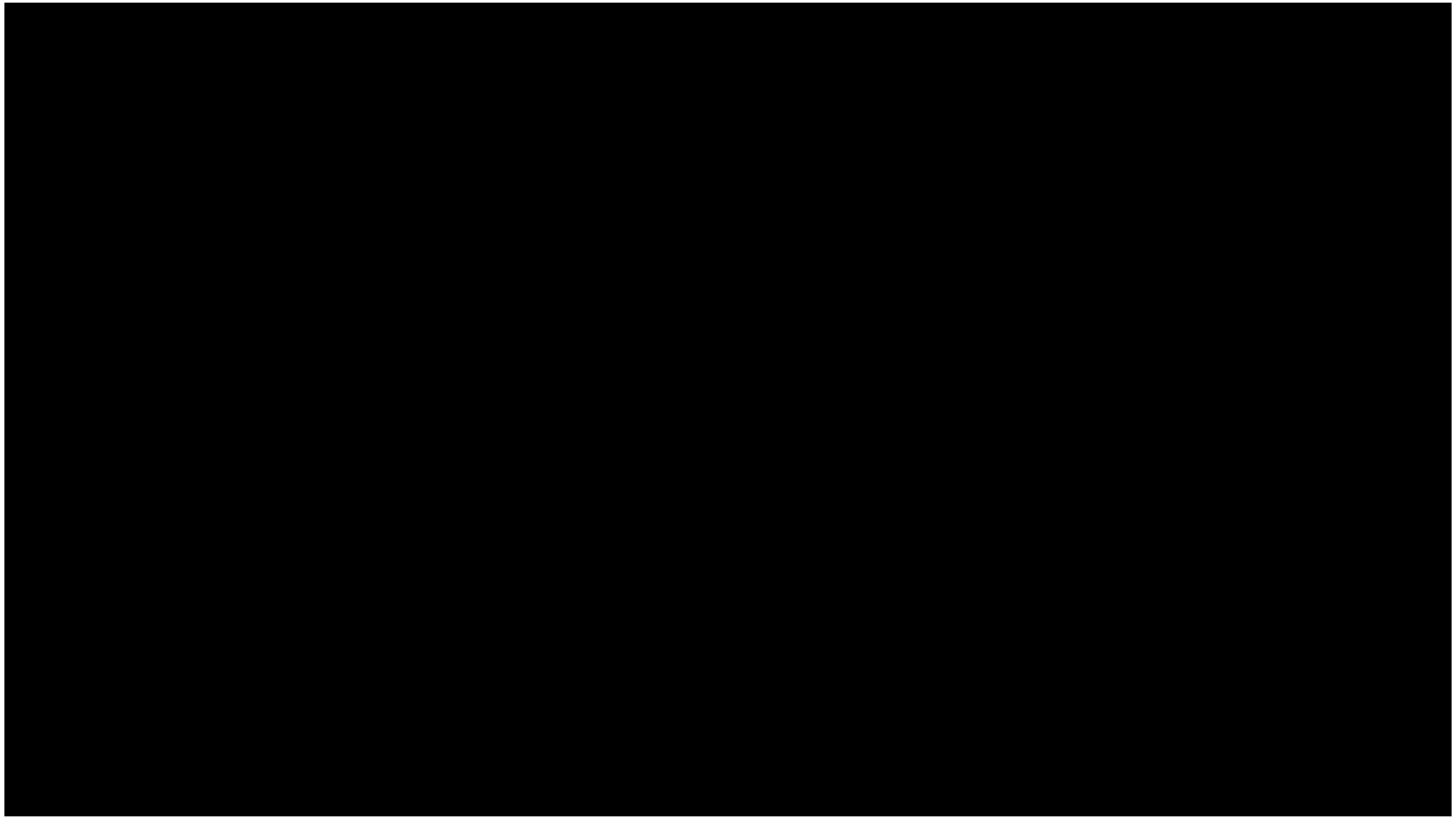
- We are modeling y with feature x



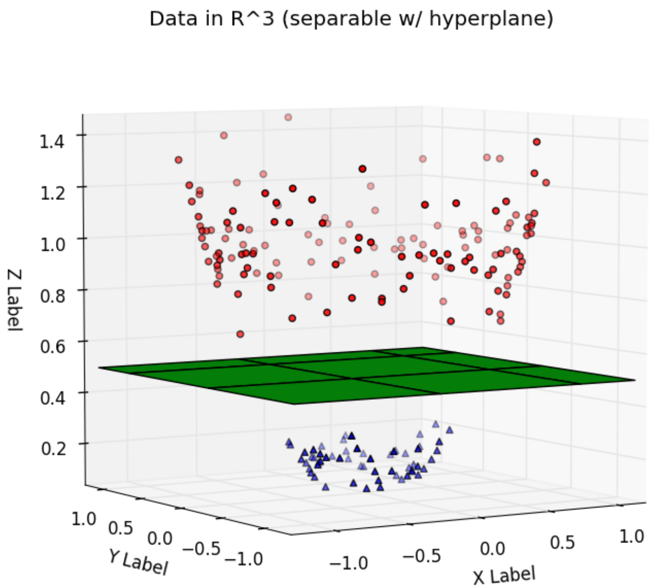
  - Classes are not separable with this feature
- One solution: non-linear classifier
- Another solution: add features!
  - E.g., x^2

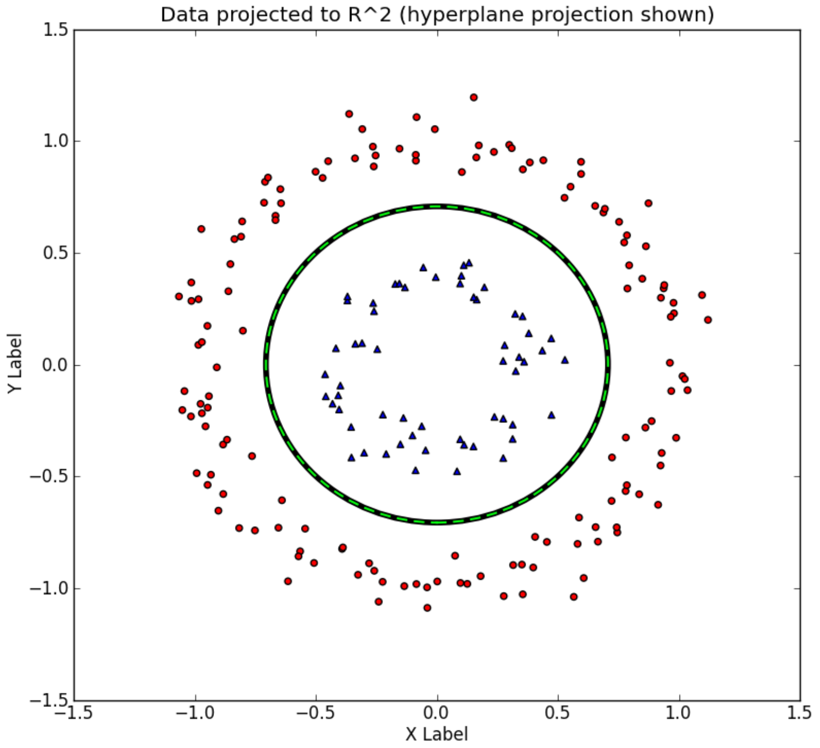


Kernel SVM



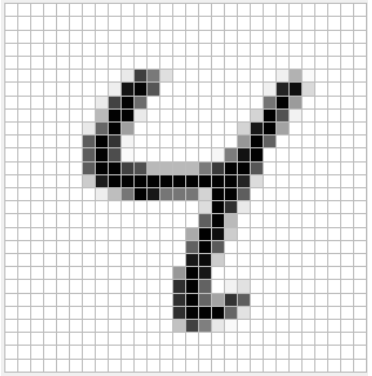
Kernel Methods: Example





Feature combinations

- Recall our feature space in digit classification
  - 28 x 28 pixels = 784 features
  - with 2nd order features: ~615k features
  - with 3rd order features: ~480m features



- Remember the "curse of dimensionality"?
  - We don't have enough data to train
- Adding interactions can help, but adding too many can hurt

Key Concepts (this lecture)

- Logistic regression
- Simplified sigmoid cost function
- Odds ratios
- Overfitting revisited
- Support vector machines
- Hard vs. soft margins
- Kernel functions