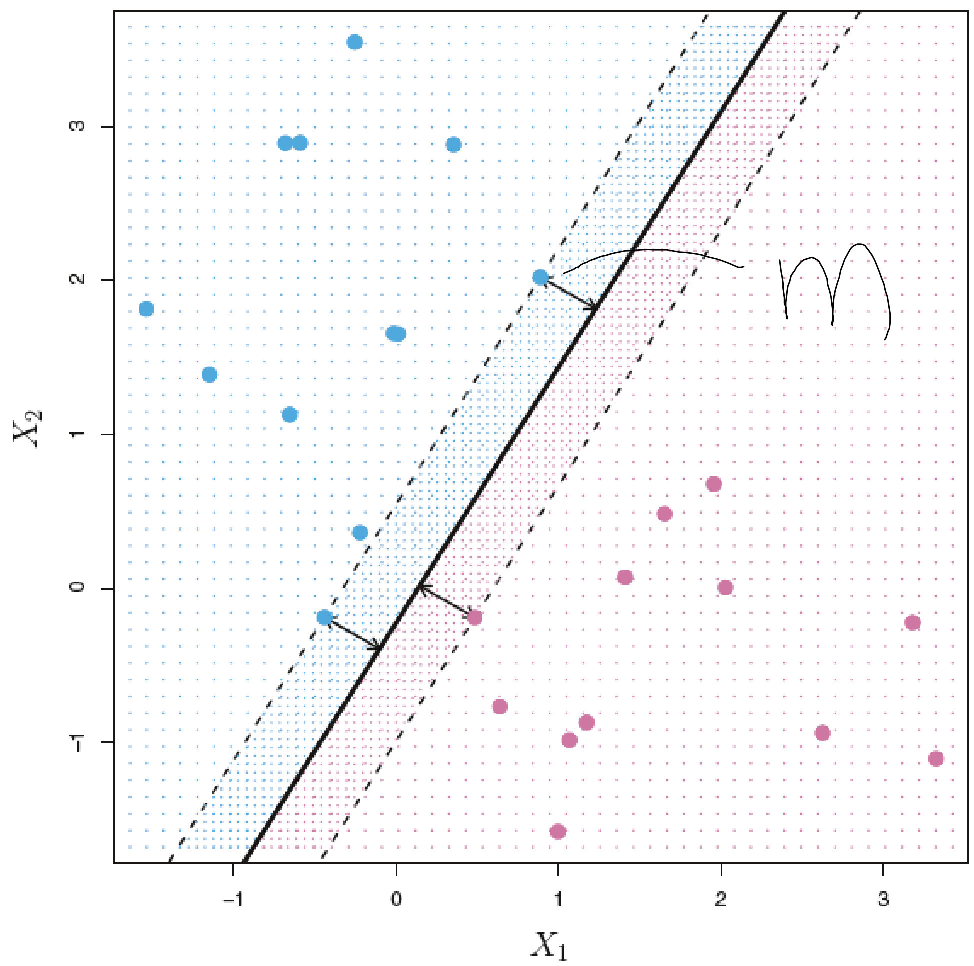
Math 407

Support Vector Machines

Maximal Marginal Classifier

Goal: Classify a binary variable Y based on the line that "maximally" separates the training set.



Some Notes on the Maximal Marginal Classifier

- Requires that the training set be separable by a "line"

- If there are p predictors, the "line" is really a "hyperplane" of dimension p-1.

Example:

If p=3, then

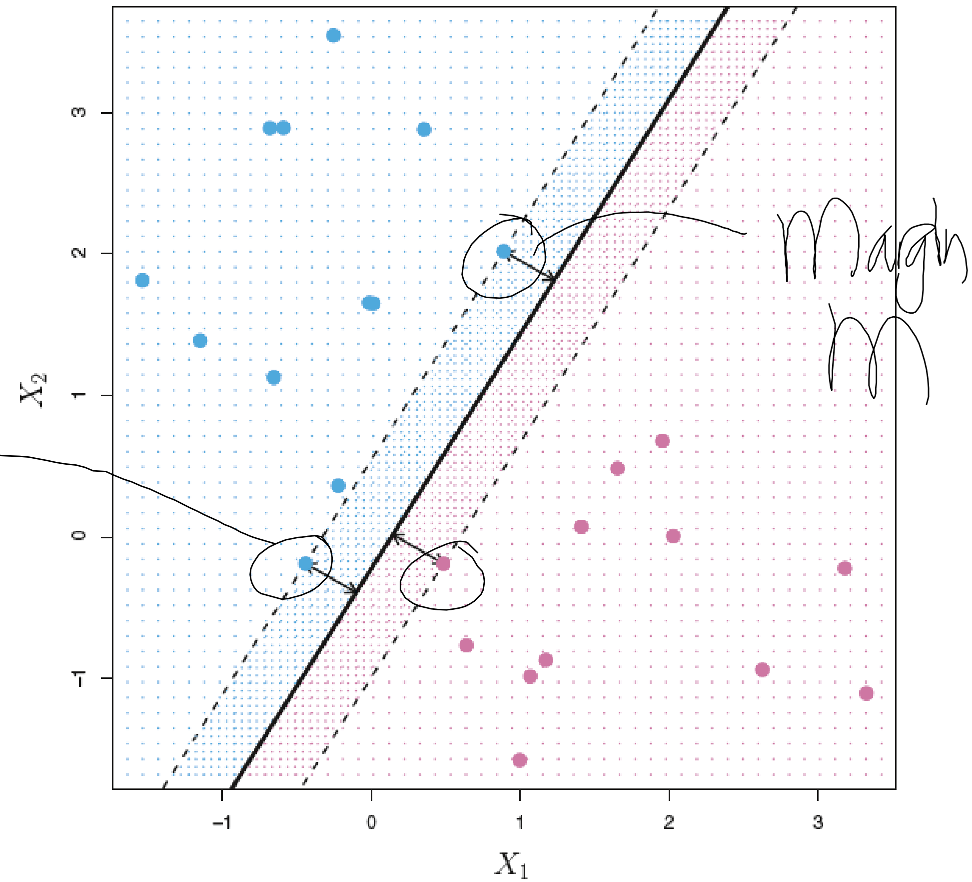
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 = 0$$

is the equation of a 2D plane.

Some Terminology

The ***margin*** is the smallest distance from the training set to the "maximally" separating line.

The ***support vectors*** are the data points in the training set that are on the margin (i.e, the closest points in the training set).



Notation

Suppose we wish to classify Y as +1 or -1 based on p predictors, $\mathbf{X}=(X_1, X_2, X_3,\ldots X_p)^T$.

Let $H(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ be a **linear** classifier where

$$\hat{Y} = +1 \text{ if } H(\mathbf{X}) > 0 \quad \text{or } -1 \text{ if } H(\mathbf{X}) < 0$$

Note that the classifier H is correct if $YH(\mathbf{X}) > 0$
and incorrect if $YH(\mathbf{X}) < 0$

How can we train a maximal margin classifier?

Given: n data points in our training set, i.e.

x_{i1}, x_{i2}, x_{i3},..., x_{ip}, and y_i for i = 1,2,...,n

Goal: find $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ so that the margin M is as large as possible

How can we train a maximal margin classifier?

Given: n data points in our training set, i.e.

$x_{i1}, x_{i2}, x_{i3},\ldots, x_{ip}$, and y_i for $i = 1,2,\ldots,n$

Goal: find $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ so that the margin M is as large as possible, i.e.

find the widest seperating interval so that all y_i's are correctly classified by

$H(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$

How can we train a maximal margin classifier?

Given: n data points in our training set, i.e.

$x_{i1}, x_{i2}, x_{i3},\ldots, x_{ip}$, and y_i for i = 1,2,...,n

Goal: find $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ so that the margin M is as large as possible, i.e.

find the widest seperating interval so that all y_i's are correctly classified by

$H(\mathbf{x_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p\, x_{ip}$

and "correctly classified" is equivalent to

$$y_i\,(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p\, x_{ip}) > 0$$

Distance from a point to a hyperplane H(x)

From vector calculus, the distance from the ith training point to H(x) is

$$y_i H(x_i) > M$$

$$\frac{|\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p\, x_{ip}|}{\sqrt{\beta_0^2 + \beta_1^2 + \beta_2^2 + \cdots + \beta_p^2}}$$

$$\sum_{i=0}^{p} \beta_i^2 = 1$$

Optimization Problem

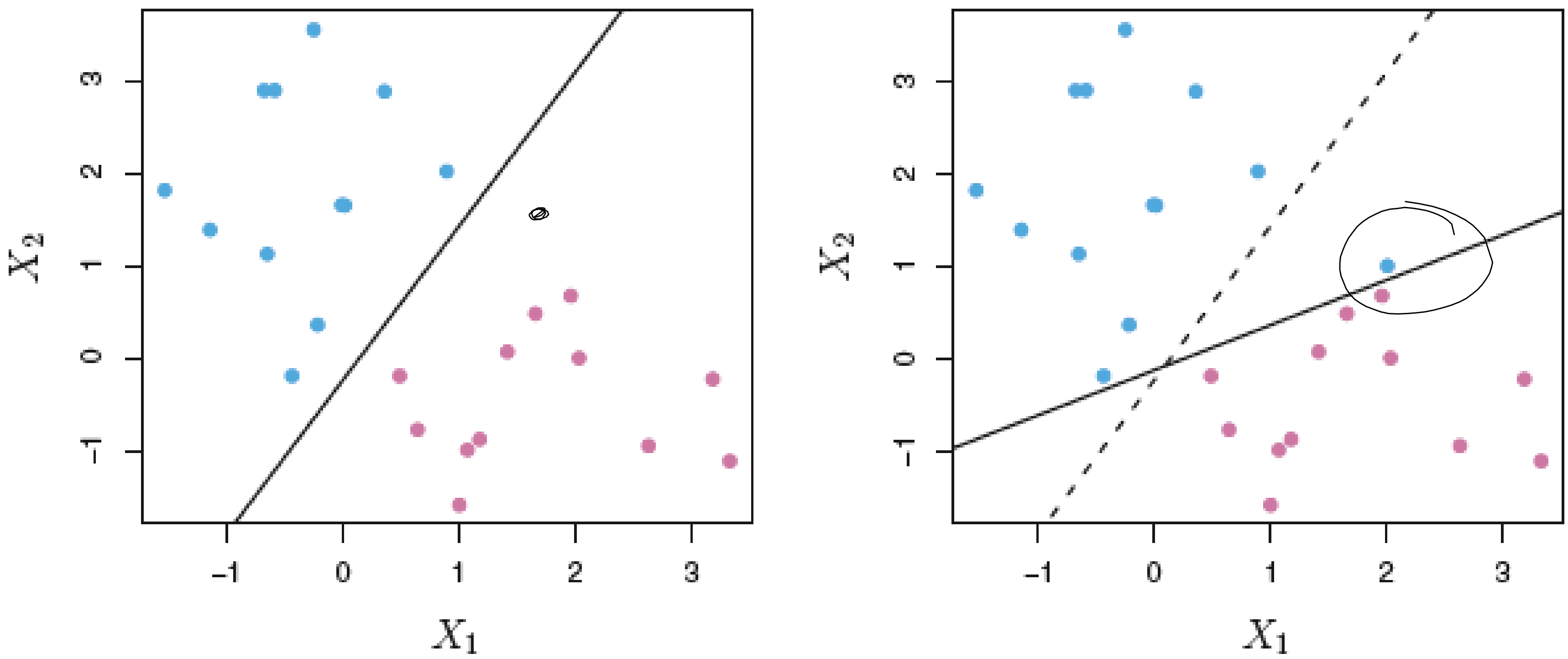
$$\underset{\beta_0,\beta_1,\ldots,\beta_p,M}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M$$

$$y_i(H(x_i))$$

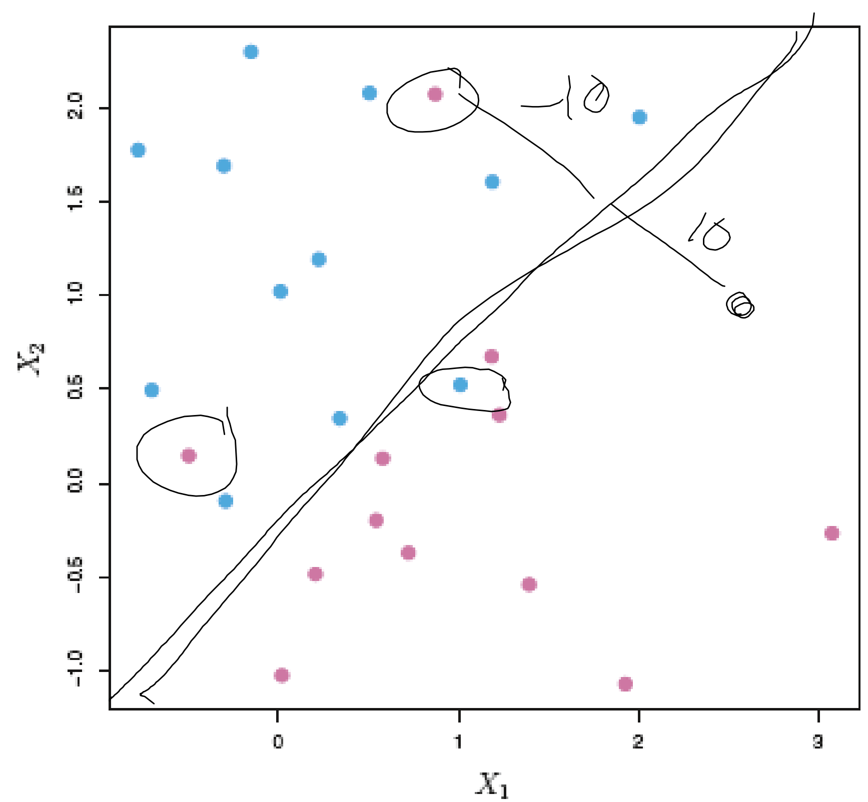
MMC may have high variance



Support Vector Classifier

Relax the assumption that the training set is separable by allowing training point to “pay” for being on the wrong side.

Training points that lie **within or on** the margin are called “support vectors”



Notation

Let ε_i be the amount the ith training point "spends" for being on the wrong side of the line.

Let C be the "budget" that the training points are allow to spend on being on the wrong side of the line.

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

Optimization Problem

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,M}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^{p}\beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1-\epsilon_i),$$

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

$\epsilon_i = 0$

$1 > \epsilon_i > 0$

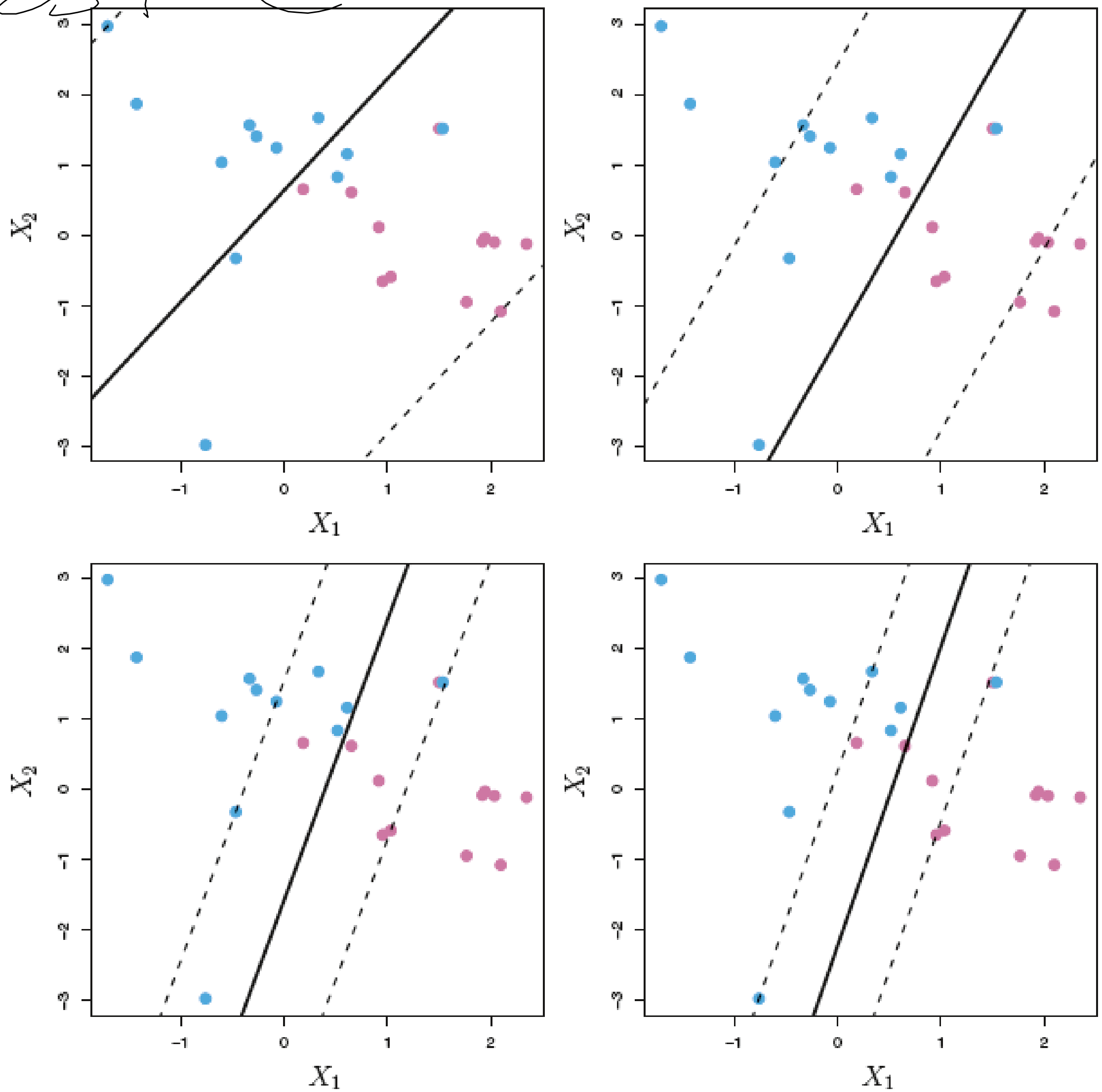
within margin, right side of line

$\epsilon_i > 1$

wrong side of line

Effect of the choice of C

largest C



smallest C

Extension to Quadratic Boundary

$$\underset{\beta_0,\beta_{11},\beta_{12}\ldots,\beta_{p1},\beta_{p2},\epsilon_1,\ldots,\epsilon_n,M}{\text{maximize}} \quad M$$

$$\text{subject to } y_i\left(\beta_0 + \sum_{j=1}^{p}\beta_{j1}x_{ij} + \sum_{j=1}^{p}\beta_{j2}x_{ij}^2\right) \geq M(1-\epsilon_i)$$

$$\sum_{i=1}^{n}\epsilon_i \leq C, \quad \epsilon_i \geq 0, \quad \sum_{j=1}^{p}\sum_{k=1}^{2}\beta_{jk}^2 = 1.$$

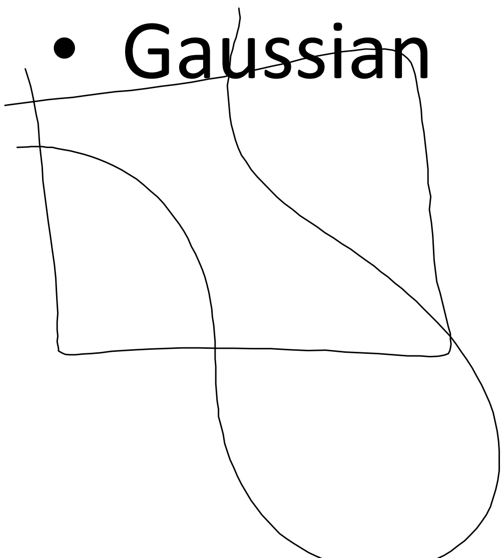
Extension to a flexible boundary

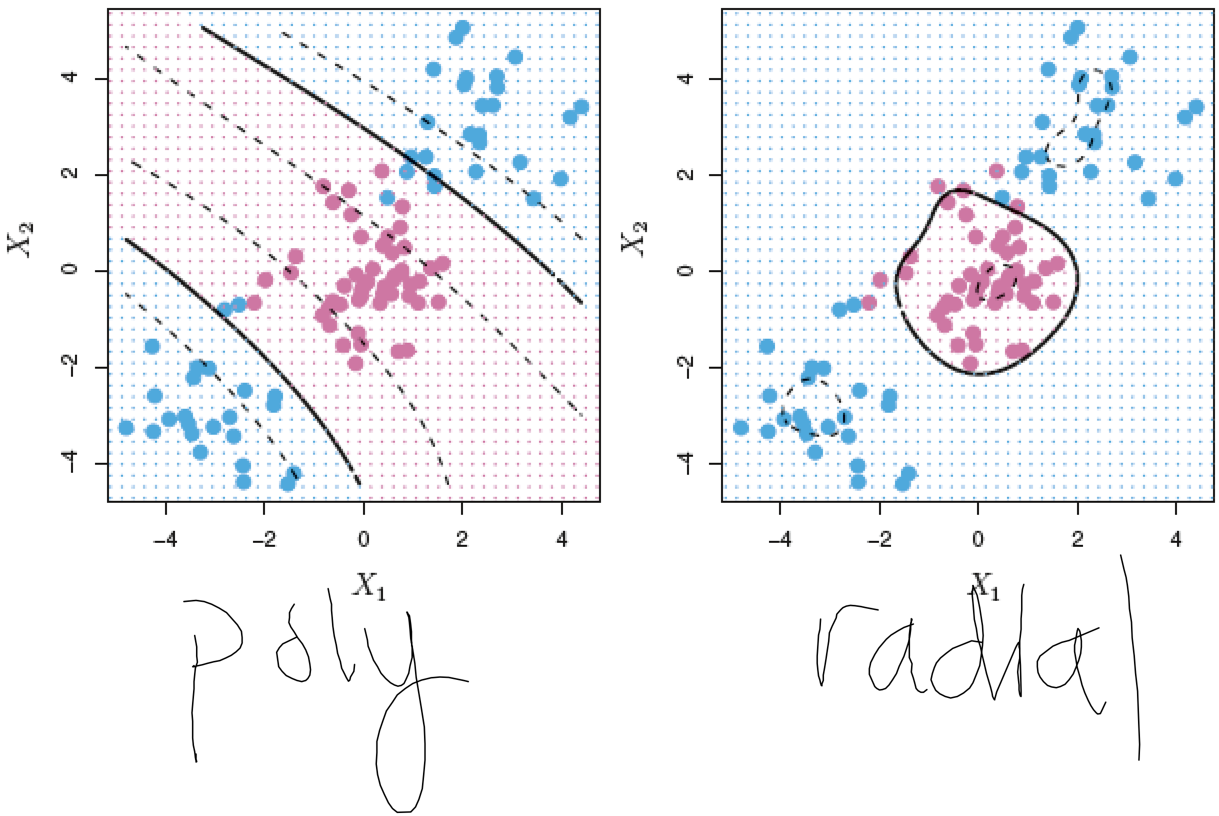
- Could use a technique like Taylor Approximation with dth degree polynomial or a penalized spline with a numerically stable set of basis functions or… use kernels.

- We briefly saw kernels when we used kernel density estimation to estimate the distribution of a predictor and then generate new observations.

Common Choices of Kernels

- Linear
- Polynomial
- Radial
- Gaussian





Support Vector **Machines** (SVMs)

a support vector classifier extended through the use of kernels to have non-linear decision boundaries.

- Choice of kernel
- Find C, the budget for misclassifications, through cross-validation

Similar performance to logistic regression, does better when points are close to separable.