LEC 6: Classification Intro and SVMs

Mar 11, 2020

Quiz

Link: https://forms.gle/MySQLbxAKGeGHKpH9

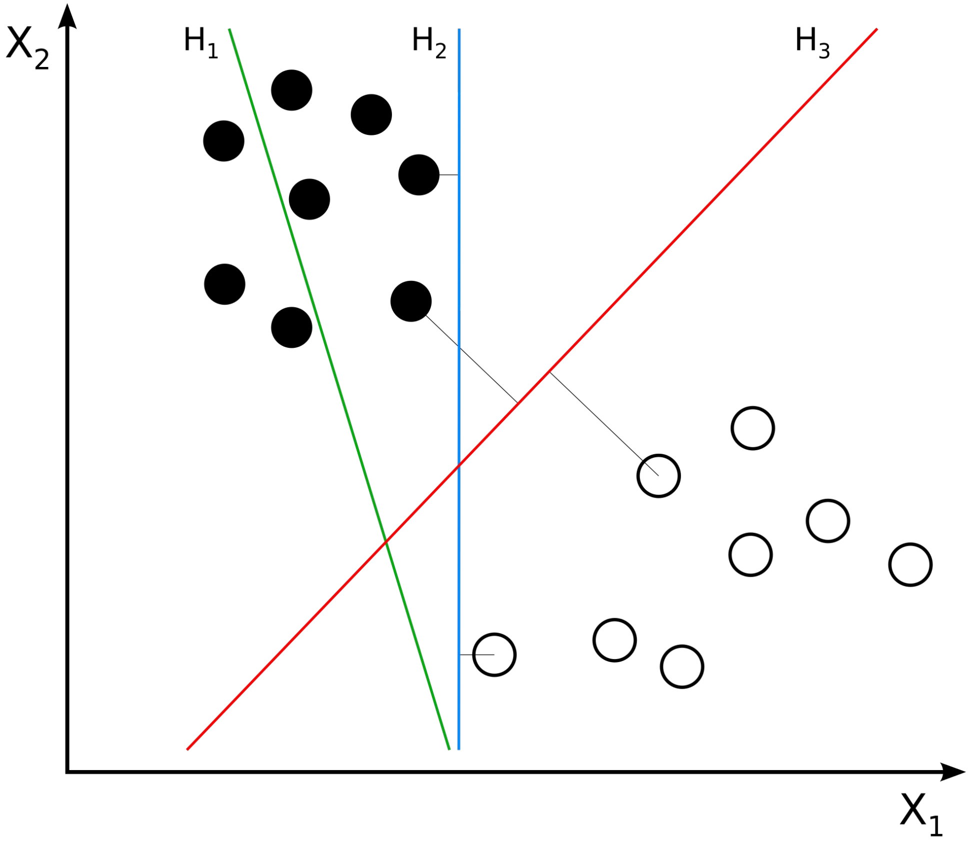
Presentation:

https://docs.google.com/presentation/d/13PTrN33nlkkQ7YDolVeCcSUSgmHj5AVGKpmfEaH0Zb0/edit?usp=sharing

Motivation for Classification

Some models for classification

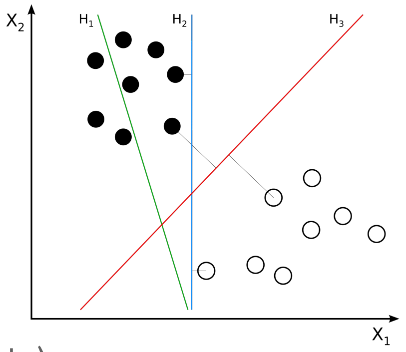
1. Supervised - training data with labels provided
    a. Logistic regression and Maximum Likelihood Estimation
    b. Support Vector Machines
    c. K-Nearest Neighbors
    d. Decision Trees and Random Forest
    e. Neural Networks
2. Unsupervised - training data does not require labels
    a. K-Means
    b. Expectation Maximization


X_2
H_1
H_2
H_3
X_1

Model: Support Vector Machines

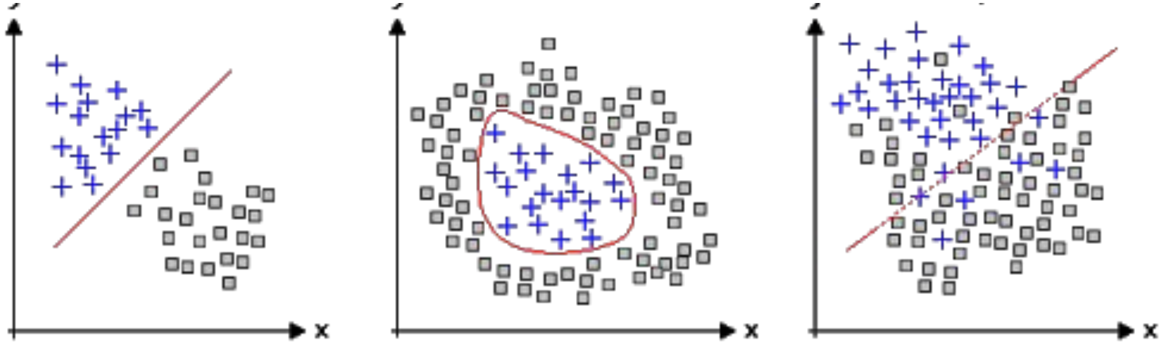
Teaching this model first to introduce idea of a Decision Boundary

- Finds linear decision boundary in N-dimensional space
  - Line in N-dimensional space is called a hyperplane.
  - Hyperplane is N-1 dimensions
- Finds the BEST linear decision boundary
  - Maximum margin - largest separation between the two classes
- Pros:
  - Low computational load for prediction (works well with high d image data)
  - Useful for supposedly linearly separable data e.g. protein classification, image feature classification
- Cons:
  - Does not give probabilistic interpretation unlike e.g. Logistic regression
  - Hyperparameter tuning is computationally expensive



Linear separability

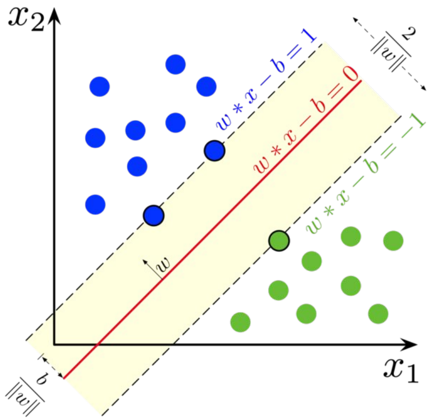
If the data can be separated by an N-1 dimension hyperplane



First, assume the data is linearly separable

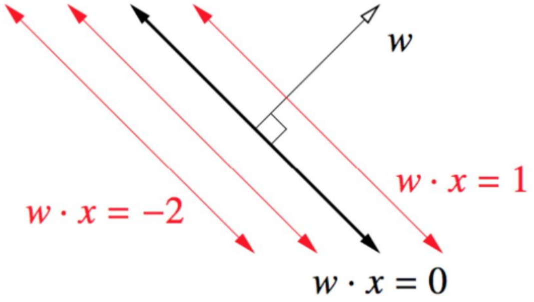
Hard margin SVMs

- **Model:** {Class 0: y = -1, Class 1: y = 1}
- **Target result:** want a function that outputs -1 when Class 0 and 1 when Class 1
- **Minimize:** Loss function (error)



Math behind SVMs - Target Result Definition

- How do we find the whether a point is above/below the boundary?
  - Use normal vector **w**
  - Dot product: how much are the vectors going in the same direction?



- Let’s try this.

Target Result Definition

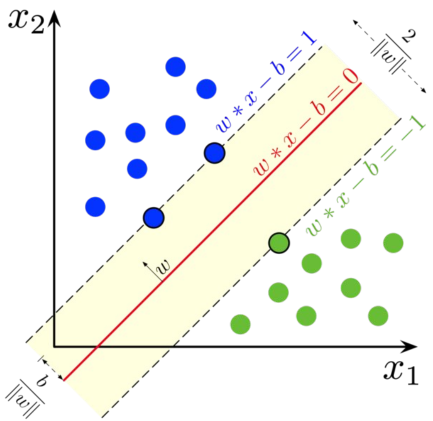
Therefore, we can define the decision boundary as

$$\forall i, \quad \begin{cases} \mathbf{w}^\top \mathbf{x}_i - b \geq 0 & \text{if } y_i = 1 \\ \mathbf{w}^\top \mathbf{x}_i - b \leq 0 & \text{if } y_i = -1 \end{cases}$$

$$\forall i, \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 0$$

- We need the -b term to maintain the offset of the hyperplane
-

Loss Function Optimization

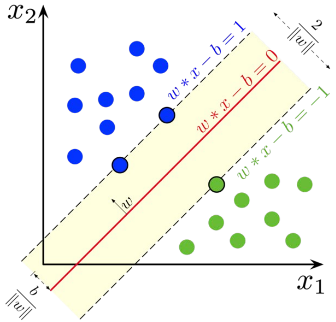


What is the metric we want to maximize?

- The width of the margin (in yellow)
- Defining the margin
  - Margin is the distance between the two lines parallel to the boundary that go through the points nearest to the boundary
  - Only need to look at points closest to decision boundary
    - Call these points “support vectors”

- Mathematically:
    - We find **w** that forces the equations to be true and derive the width of margin as line 3
    - Note the euclidean distance calculation using dot product

$$w^T \cdot x_+ - b = +1$$
$$w^T \cdot x_- - b = -1$$
$$\frac{w}{||w||} \cdot (x_+ - x_-) = \frac{w^T \cdot (x_+ - x_-)}{||w||} = \frac{2}{||w||}$$



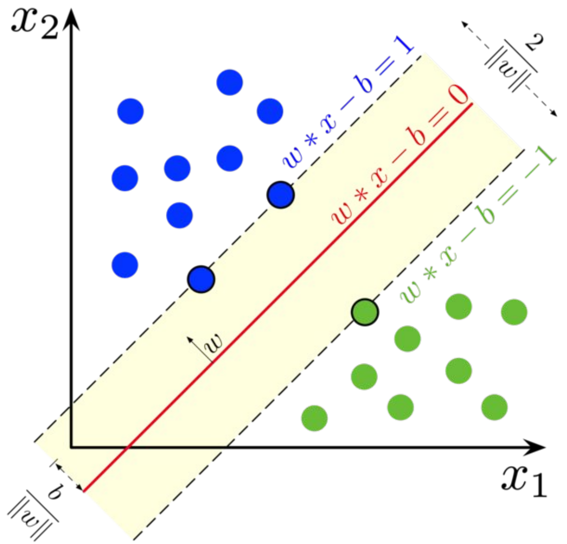
- Therefore maximize $\frac{2}{||w||}$ such that we maintain the constraint
    - Constraint that for all points, they are correctly categorized
    - Same as saying minimize $||w||$

$$\forall i, \quad \begin{cases} \mathbf{w}^\top \mathbf{x}_i - b \geq 0 & \text{if } y_i = 1 \\ \mathbf{w}^\top \mathbf{x}_i - b \leq 0 & \text{if } y_i = -1 \end{cases}$$

$$\forall i, \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 0$$

Calculate decision boundary to get classifier

- After finding w, boundary is w*x - b = 0

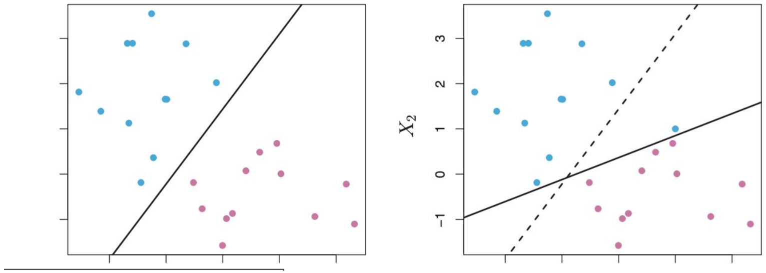


Constraint of Hard Margin SVMs

Hard margin SVMs problems: must be linearly separable, too sensitive to outliers

$$\forall i, \quad \begin{cases} \mathbf{w}^\top \mathbf{x}_i - b \geq 0 & \text{if } y_i = 1 \\ \mathbf{w}^\top \mathbf{x}_i - b \leq 0 & \text{if } y_i = -1 \end{cases}$$

$$\forall i, \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 0$$



Soft Margin SVMs

Instead of having a hard constraint of

$$\forall i, \quad \begin{cases} \mathbf{w}^\top \mathbf{x}_i - b \geq 0 & \text{if } y_i = 1 \\ \mathbf{w}^\top \mathbf{x}_i - b \leq 0 & \text{if } y_i = -1 \end{cases}$$

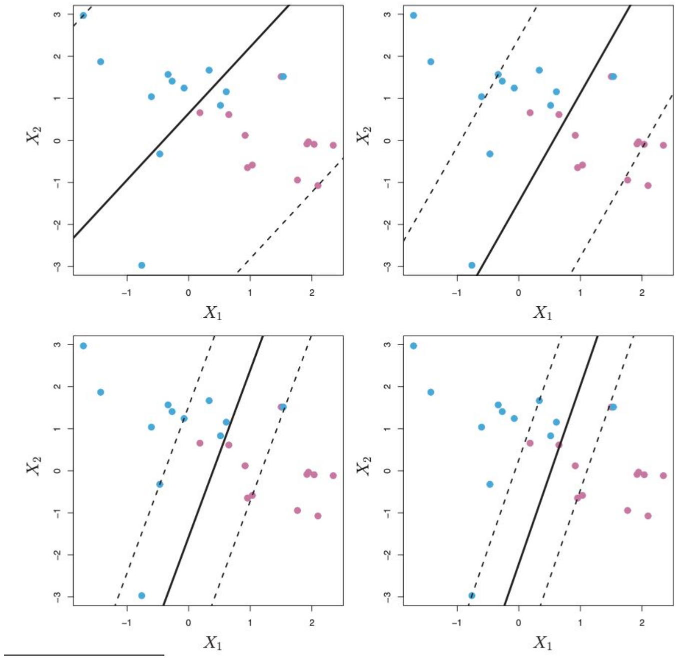
$$\forall i, \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 0$$

We define a Loss function (aka error function) that we seek to minimize with respect to **w** that incorporates the constraint

- This function is zero if the constraint is satisfied, in other words, if **x_i** lies on the correct side of the margin. For data on the wrong side of the margin, the function's value is proportional to the distance from the margin
- Use validation to choose Lambda (what does Lambda do?)

$$\left[\frac{1}{n}\sum_{i=1}^{n} \max\left(0, 1 - y_i(\vec{w} \cdot \vec{x}_i - b)\right)\right] + \lambda \|\vec{w}\|^2,$$

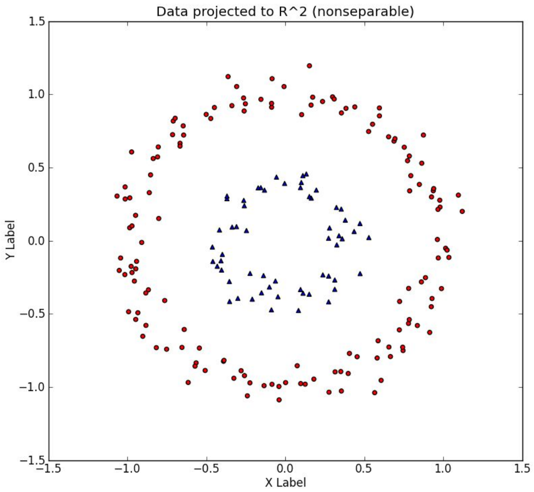
Soft margin SVMs with increasing Lambda values



What if the boundary should not be linear?

Intuition: what do we want the boundary to be?

- Lifting data into more dimensions can yield Separable data
- Add more dimensions into the **x** and **w** vectors!

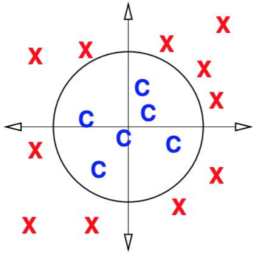


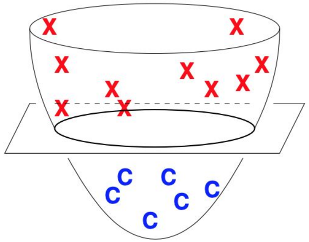
Kernels

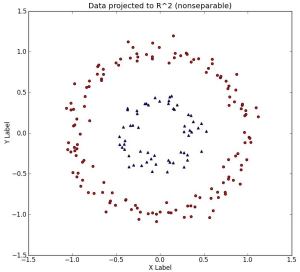
New feature is just a function of other input features

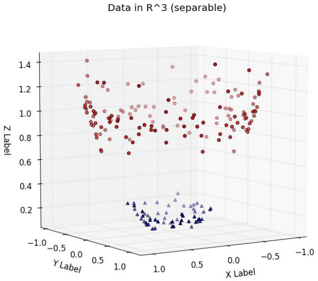
New boundary is linear in 2d

$$\begin{bmatrix} x_1 \\ x_2 \\ (x_1^2 + x_2^2) \end{bmatrix}$$





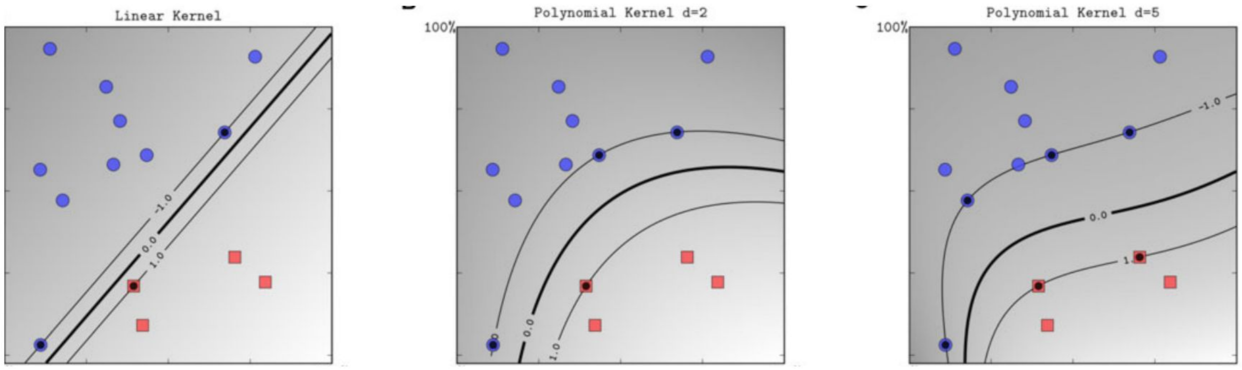




Polynomial kernel

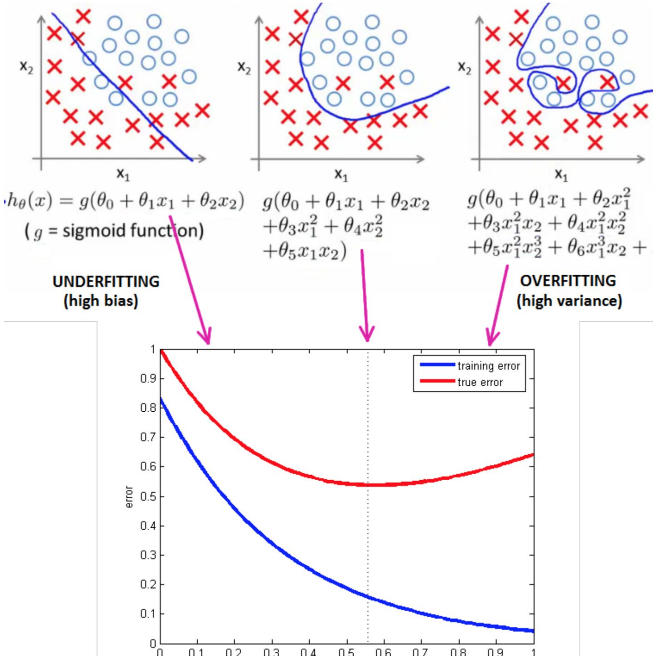
linear x = $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

Quadratic x = $\begin{bmatrix} x_1 \\ x_2 \\ x_1 * x_2 \\ x_1^2 \\ x2^2 \end{bmatrix}$



What would the cubic kernel look like?

Beware of overfitting with high degree polynomial kernels



Feedback

https://forms.gle/Uv3YfeGejQqnFXv39

https://tinyurl.com/tw7u8nd