Statistical Methods for NLP

Text Categorization, Support Vector Machines

Sameer Maskey

Announcement

- Reading Assignments
  - Will be posted online tonight
- Homework 1
  - Assigned and available from the course website
  - Due in 2 Weeks (Feb 16, 4pm)
  - 2 programming assignments

Project Proposals

- Reminder to think about projects
- Proposals due in 3 weeks (Feb 23)

Topics for Today

- Naïve Bayes Classifier for Text
- Smoothing
- Support Vector Machines
- Paper review session

Naïve Bayes Classifier for Text

$$P(y_k, X_1, X_2, ..., X_N) = P(y_k)\Pi_i P(X_i | y_k)$$

Prior Probability of the Class

Conditional Probability of feature given the Class

Here N is the number of words, not to confuse with the total vocabulary size

Naïve Bayes Classifier for Text

$$P(y = y_k | X_1, X_2, ..., X_N) = \frac{P(y = y_k) P(X_1, X_2, .., X_N | y = y_k)}{\sum_j P(y = y_j) P(X_1, X_2, .., X_N | y = y_j)}$$

$$= \frac{P(y = y_k) \Pi_i P(X_i | y = y_k)}{\sum_j P(y = y_j) \Pi_i P(X_i | y = y_j)}$$

$$y \leftarrow argmax_{y_k} P(y = y_k) \Pi_i P(X_i | y = y_k)$$

Naïve Bayes Classifier for Text

- Given the training data what are the parameters to be estimated?

$$P(y) \qquad P(X|y_1) \qquad P(X|y_2)$$

$P(y)$	$P(X\|y_1)$	$P(X\|y_2)$
Diabetes : 0.8 Hepatitis : 0.2	the: 0.001 diabetic : 0.02 blood : 0.0015 sugar : 0.02 weight : 0.018 …	the: 0.001 diabetic : 0.0001 water : 0.0118 fever : 0.01 weight : 0.008 …

$$y \leftarrow argmax_{y_k} P(y = y_k)\Pi_i P(X_i|y = y_k)$$

Estimating Parameters

- Maximum Likelihood Estimates
  - Relative Frequency Counts
- For a new document
  - Find which one gives higher posterior probability
    - Log ratio
    - Thresholding
- Classify accordingly

Smoothing

- MLE for Naïve Bayes (relative frequency counts) may not generalize well
  - Zero counts

- Smoothing
  - With less evidence, believe in prior more
  - With more evidence, believe in data more

Laplace Smoothing

- Assume we have one more count for each element
- Zero counts become 1

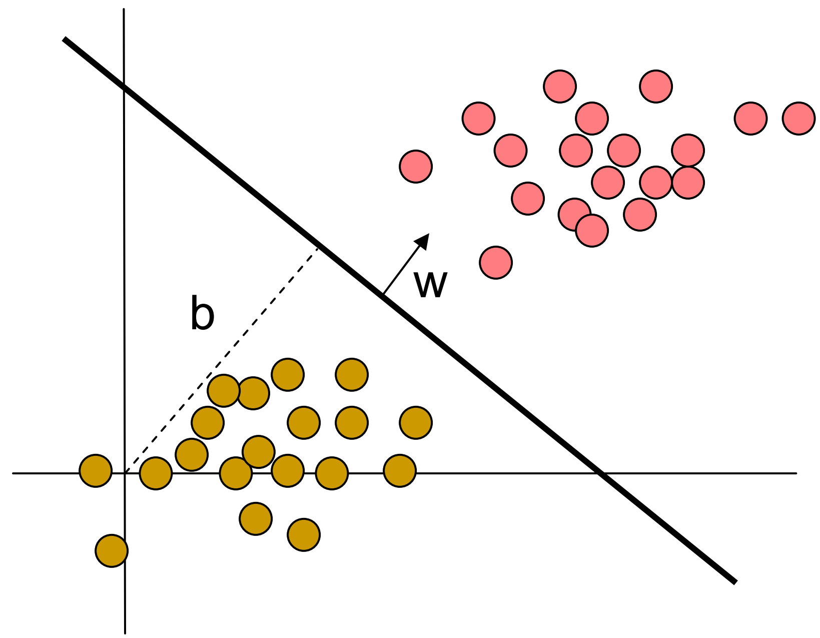
$$P_{smooth}(w) = \frac{c_w + 1}{\sum_w \{c(w) + 1\}}$$

$$P_{smooth}(w) = \frac{c_w + 1}{N + V}$$

Vocab Size

Back to Discriminative Classification

$$f(x) = \mathbf{w}^T x + b$$



Linear Classification

- If we have linearly separable data we can find w such that

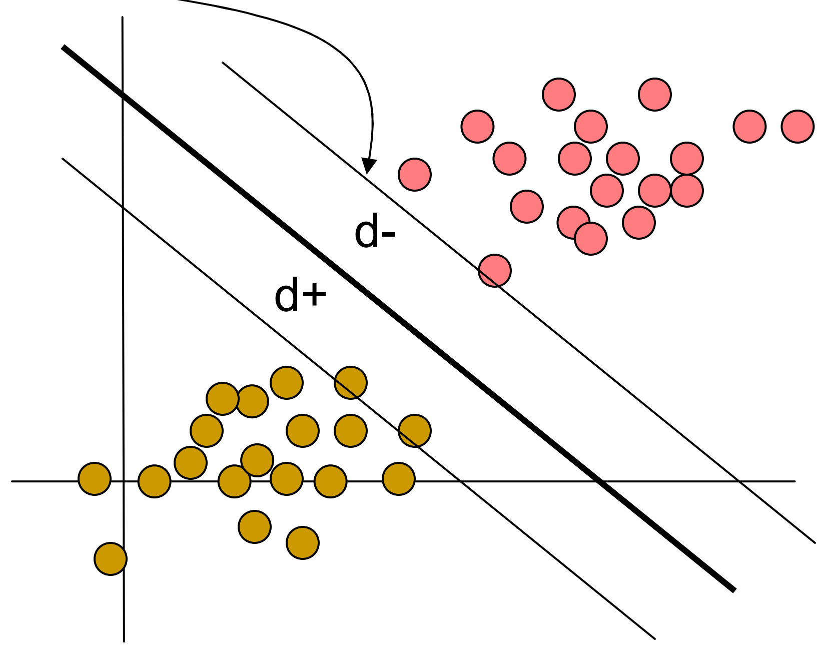
$$y_i(\mathbf{w}^T x_i + b) > 0 \ \forall i$$

Margin

- Let us have hyperplanes such that

$$\mathbf{w}^T x_i + b \geq +1 \text{ if } y_i = +1$$
$$\mathbf{w}^T x_i + b \leq -1 \text{ if } y_i = -1$$

$$y_i(\mathbf{w}^T x_i + b) - 1 \geq 0 \ \forall i$$



Total margin is sum of d+ and d-

Maximizing Margin

- Distance between H and H+ is $\frac{1}{\|w\|}$

- Distance between H+ and H- is $\frac{2}{\|w\|}$

- In order to maximize the margin need to minimize the denominator $\frac{1}{2}\|w\|^2$

Maximizing Margin with Constraints

- We can combine the two inequalities to get

$$y_i(\mathbf{w}^T x_i + b) - 1 \geq 0 \ \forall i$$

- Problem formulation
  - Minimize $\frac{\|w\|^2}{2}$
  - Subject to $y_i(\mathbf{w}^T x_i + b) - 1 \geq 0 \ \forall i$

Solving with Lagrange Multipliers

- Solve by introducing Lagrange Multipliers for the constraints
- Minimize

$$J(\mathbf{w}, b, \alpha) = \frac{\|w\|^2}{2} - \sum_{i=1}^{n} \alpha_i \{y_i(\mathbf{w}^T x_i + b) - 1\}$$

For given α_i

$$\frac{\partial}{\partial \mathbf{w}} J(\mathbf{w}, b, \alpha) = \mathbf{w} - \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial}{\partial b} J(\mathbf{w}, b, \alpha) = -\sum_{i=1}^{n} \alpha_i y_i$$

Dual Problem

- Solve dual problem instead
- Maximize

$$J(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i . x_j)$$

- subject to constraints of

$$\alpha_i \geq 0 \ \forall i$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

Quadratic Programming Problem

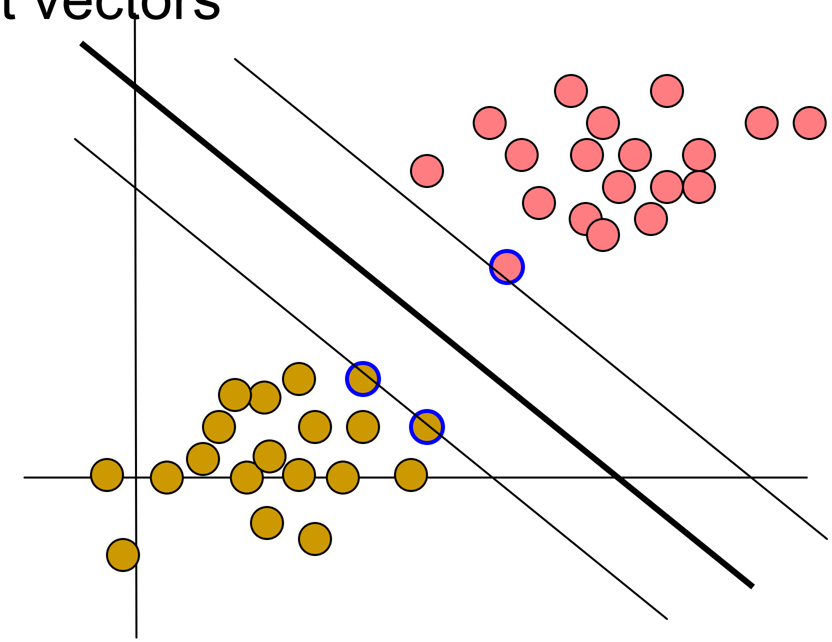
- Minimize f(x) such that g(x) = k
  - Where f(x) is quadratic and g(x) are linear constraints
- Constrained optimization problem
- Saw the example before

SVM Solution

$$\hat{\mathbf{w}} = \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i$$

- Linear combination of weighted training example
- Sparse Solution, why?
  - Weights zero for non-support vectors

$$\sum_{i \in SV} \widehat{\alpha}_i y_i (x_i . x) + \widehat{b}$$



Sequential Minimal Optimization (SMO) Algorithm

- The weights are just linear combinations of training vectors weighted with alphas
- We still have not answered how do we get alphas
  - Coordinate ascent

Do until converged

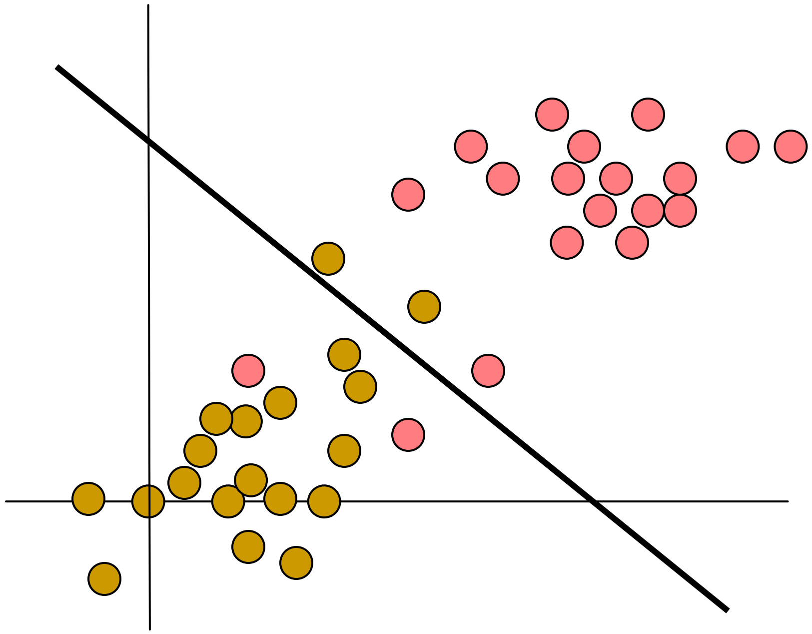
        select pair of alpha(i) and alpha(j)

        reoptimize W(alpha) with respect to alpha(i) and alpha(j)

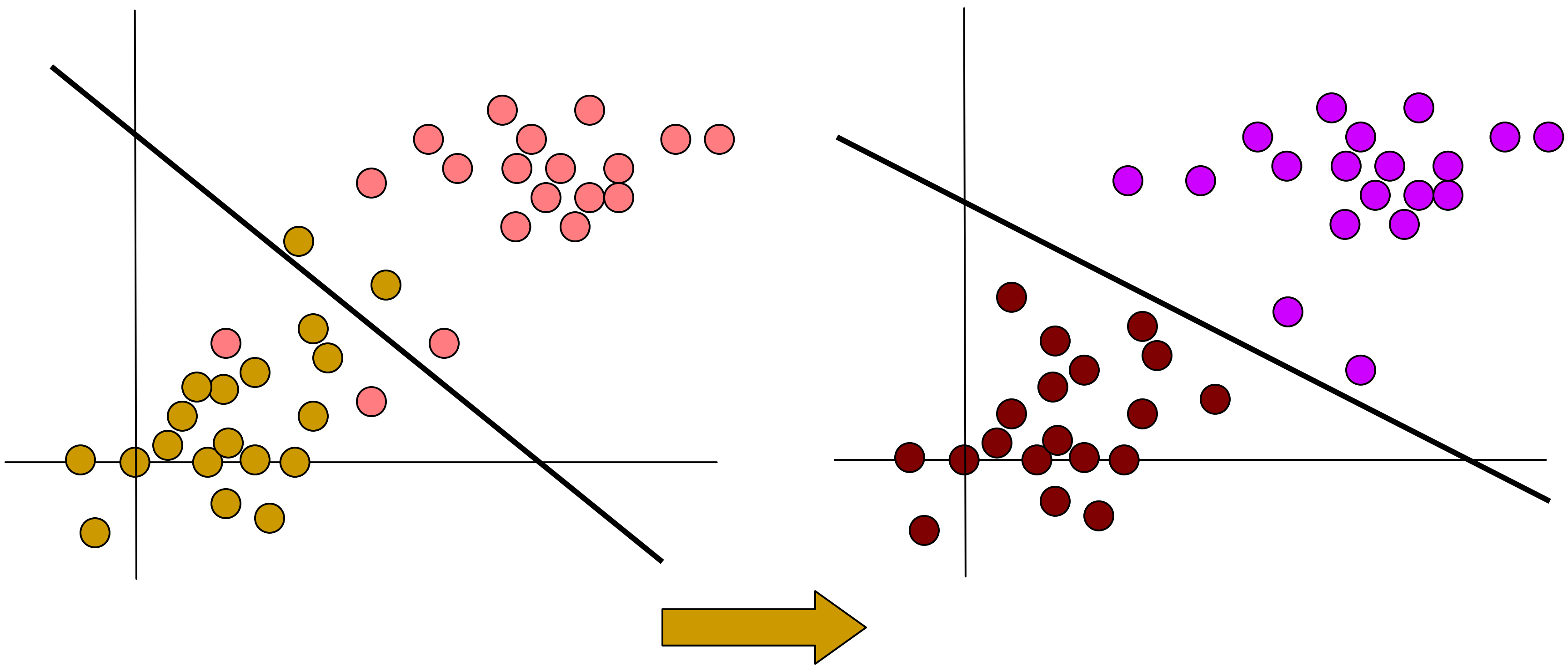
        holding all other alphas constant

done

Not Linearly Separable



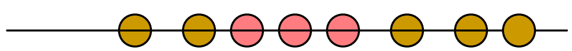
Transformation



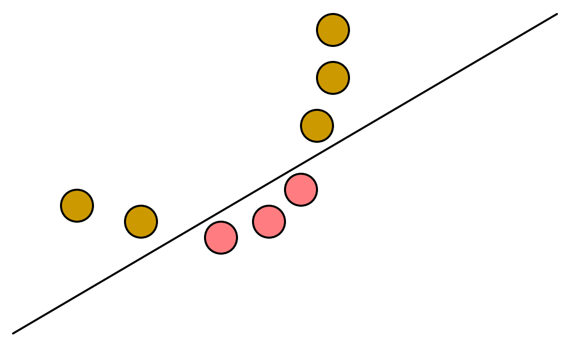
Transformation h( ) =

Non Linear SVMs

- Map data to a higher dimension where linear separation is possible
- We can get a longer feature vector by adding dimensions



$$x \rightarrow (x^2, x)$$



$$\phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1)$$

Kernels

Given feature mapping $\phi(x)$ define
$K(x, z) = \phi(x)^T \phi(z)$

$$\phi(x)^T \phi(z)$$

$$= x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1 x_2 z_1 z_2 + 2x_1 z_1 + 2x_2 z_2 + 1$$

$$= (x.z + 1)^2$$

May not need to explicitly transform

Example of Kernel Functions

$$K(x,z) = x.z$$

Linear Kernel

$$K(x,z) = (x.z+1)^p$$

Polynomial Kernel

$$K(x,z) = exp(-\frac{\|x-z\|^2}{2\sigma^2})$$

Gaussian Kernel

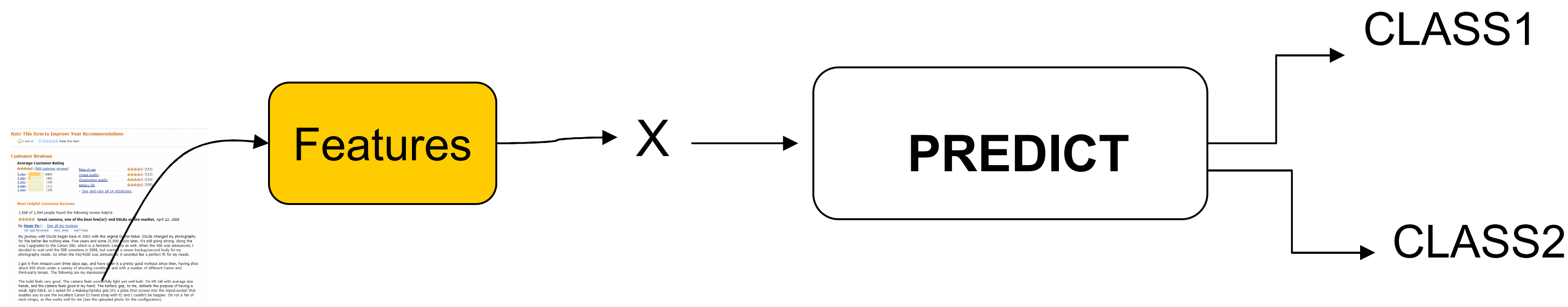
Non-separable case

- Some data sets may not be linearly separable
- Introduce slack variable
- Also helps regularization
  - Less sensitive to outliers

  - Minimize $\frac{\|w\|^2}{2} + C\sum_{i=1}^{n}\xi_i$
  - Subject to $y_i(\mathbf{w}^T x_i + b) \geq 1 - \xi_i \;\forall i$

    $\xi_i \geq 0 \;\forall i$

Summary



- Linear Classification Methods
  - Fisher's Linear Discriminant
  - Perceptron
  - Support Vector Machines

References

- Tutorials on www.svms.org/tutorial