Lecture 12 Support Vector Machine II

ECE 625: Data Analysis and Knowledge Discovery

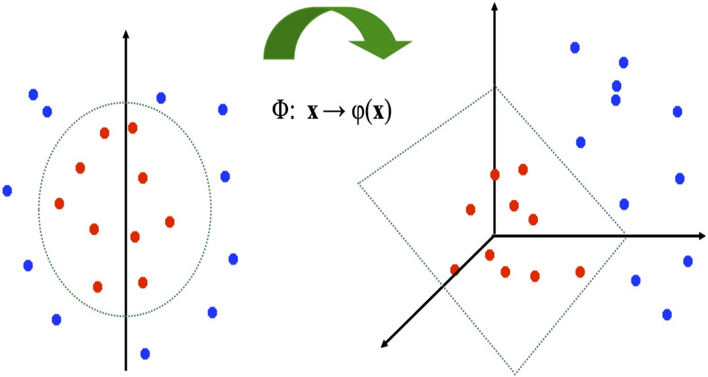
Di Niu

Department of Electrical and Computer Engineering
University of Alberta

March 4, 2021

Outline

Feature Expansion



Feature Expansion

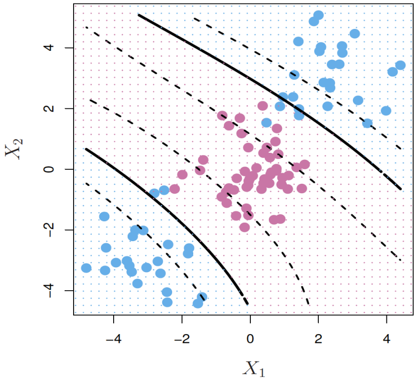
- Enlarge the space of features by including transformations; for example $X_1^2, X_2^3, X_1X_2, X_1X_2^2, \cdots$, Hence go from a p-dimensional space to an $M > p$ dimensional space.
- Fit a support-vector classifier in the enlarged space.
- This results in non-linear decision boundaries in the original space.
- Example: Suppose we use $(X_1, X_2, X_1^2, X_2^2, X_1X_2)$ instead of just (X_1, X_2). Then the decision boundary would be of the form

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_5 X_1 X_2 = 0.$$

- This leads to nonlinear decision boundaries in the original space (quadratic conic sections).

Cubic Polynomials

- Here we use a basis expansion of cubic polynomials — from 2 variables to 9.
- The support vector classifier in the enlarged space solves the nonlinear classification problem in the original lower-dimensional space



- The decision boundary is

$$\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_1^2+\beta_4X_2^2+\beta_5X_1X_2+\beta_6X_1^3+\beta_7X_2^3+\beta_8X_1X_2^2+\beta_9X_1^2X_2 = 0.$$

Nonlinearities and Kernels

- Polynomials (especially high-dimensional ones) get wild rather fast.
- There is a more elegant and controlled way to introduce nonlinearities in support vector classifier — through the use of kernels.
- Before we discuss these, we must understand the role of inner products in support vector classifier.

Inner products and kernels

- Inner product between vectors

$$\langle x_i, x_{i'} \rangle = \sum_j x_{ij} x_{i'j}.$$

x = (x1,…,xp)

- In theory [Sec. 12.2.1], the linear support vector classifier can be represented as

$$f(x) = \beta_0 + \sum_i \alpha_i \langle x, x_i \rangle$$ =0

x_i are training samples

- To estimate parameters $\alpha_1, \cdots, \alpha_n$ and β_0, we need all $\binom{n}{2}$ inner products $\langle x_i, x_j \rangle$ between all pairs of training samples.
- It turns out that most of the $\hat{\alpha}_i$ can be zero

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \hat{\alpha}_i \langle x, x_i \rangle,$$

If xi is a support vector, alpha_i is nonzero; otherwise it's zero

where $\mathcal{S}$ is the support set of indices i such that $\hat{\alpha}_i > 0$.

Kernels and Support Vector Machine

f(x) is determined by only the support vectors.

- If we can compute inner products between observations, we can fit a support vector classifier — can be very abstract!
- Some special kernel function can do this for us. E.g.

$$K(x_i, x_{i'}) = (1 + \sum_j x_{ij} x_{i'j})^2$$

computes the inner products needed for d dimensional polynomials.

- The solution has the form

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \hat{\alpha}_i K(x, x_i).$$

Radial Kernel

- The radial Kernel has the format

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_j (x_{ij} - x_{i'j})^2\right),$$

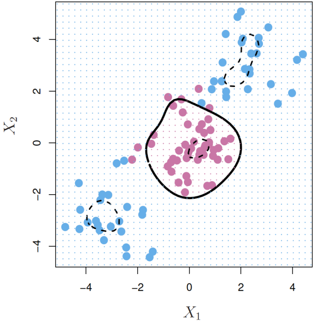
where γ is tuning parameter.

- The decision bounady is,

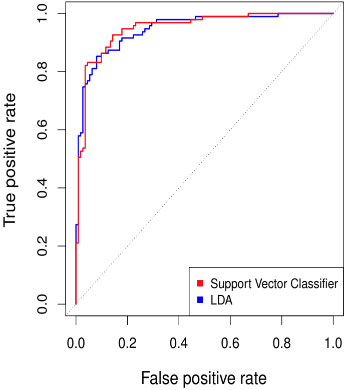
$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \hat{\alpha}_i K(x, x_i),$$

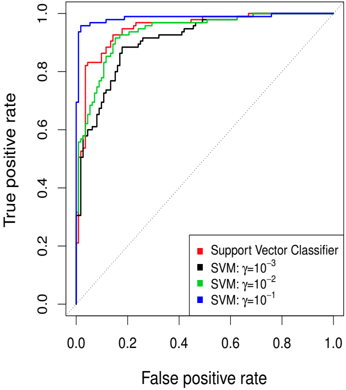
the implicit feature space is very high dimensional.

- But we just need to compute kernels for all pairs of observations.



Example - Heart Data



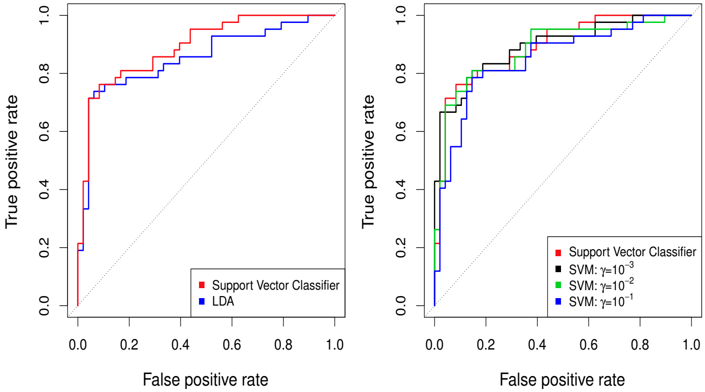


Receive Operating Curve

ROC curves on Training data

- ROC curve is obtained by changing the threshold 0 to threshold t in $\hat{f}(X) > t$, and recording false positive and true positive rates as t varies.

Example - Heart Data



ROC curves on Testing data

SVMs: More than 2 classes

- The SVM as defined works for $K = 2$ classes. What do we do if we have $K > 2$ classes?
- OVA - One versus All. Fit K different 2-class SVM classifiers $\hat{f}_k(x),\ k = 1, \cdots, K$; each class versus the rest. Classify x^* to the class for which $\hat{f}_k(x^*)$ is largest.
- OVO - One versus One. Fit all $\binom{K}{2}$ pairwise classifiers $\hat{f}_{kl}(x)$. Classify x^* to the class that wins the most pairwise competitions.
- Which one to choose? If K is not too large, use OVO.

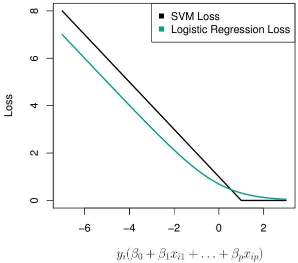
Support Vector Machine Versus Logistic Regression

- Let $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, it is found that support vector machine can be rephrased as

$$\text{minimize}_{\beta_0,\beta_1,\cdots,\beta_p} \left\{ \sum_i \max[0, 1 - y_i f(x_i)] + \lambda \sum_j \beta_j^2 \right\},$$

where γ is tuning parameter.

- This has the form of loss plus penalty.
- The loss is known as hinge loss.
- Very similar to the loss in logistic regression (negative log-likelihood).



Kernels and Support Vector Machine

- When classes are (nearly) separable, SVM does better than LR, which does better than LDA.
- When not, LR (with ridge penalty) and SVM very similar.
- If you wish to estimate probabilities, LR is the choice.
- For nonlinear boundaries, kernel SVMs are popular. Can use kernels with LR and LDA as well, but computations are more expensive.

Summary and Remark

- Feature expansion
- Kernel trick
- More than 2 classes
- Read textbook Chapter 12 and R code
- Do R lab