Exploration of classification methods: SVM and KDE

Xi Cheng, Heng Xu, Jing Peng, Zimeng Wang, Andy Wu, Shiyuan Li

University of California, Davis

Instructor: Xiaodong Li

RTG

June 2017

Project Summary

- **Goal:** To publish a Wiki page and draft text notes detailing the classification methods Support Vector Machines (SVM) and Kernel Density Classification (KDC) so that anyone may learn about them
- **Part 1: Conceptual Study**
- **Part 2: Empirical Analysis**

What is Classification?

- Classification is the problem of identifying which category a new observation belongs to, given a set of features for that observation and a set of observations whose category is known
- Example: Classifying email into spam vs. non-spam

Hard Margin

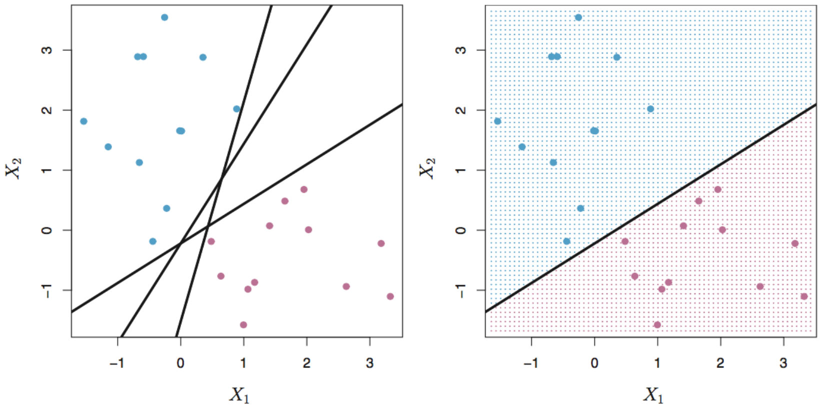
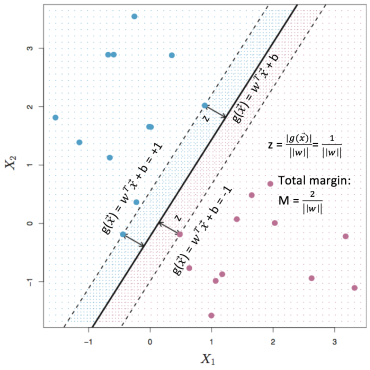


Figure: Infinitely many hyperplanes; hyperplane separates data into 2 classes; Our goal is to use training data to develop a classifier to correctly classify test data with certain constraints

Maximal Margin Classifier

Optimal separating hyperplane: hyperplane that has the farthest minimum distance to the training observations.



Primal Optimization Problem:

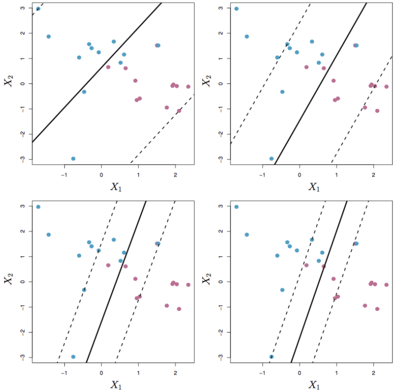
$$\underset{w,b}{\text{maximize}} \quad \frac{2}{||w||}$$

$$s.t. \quad y_i(w^T x_i + b) \geq 1,$$

$$i = 1, \ldots, n$$

Support Vector Classifier

Described by a soft margin, allowing some observations to be on the wrong side of the margin or even incorrect side of the hyperplane subject to a cost parameter



Primal Optimization Problem:

$$\underset{w,b,\epsilon_i}{\text{minimize}} \quad \frac{1}{2}w^T w + C\sum_{i=1}^{n}\epsilon_i$$

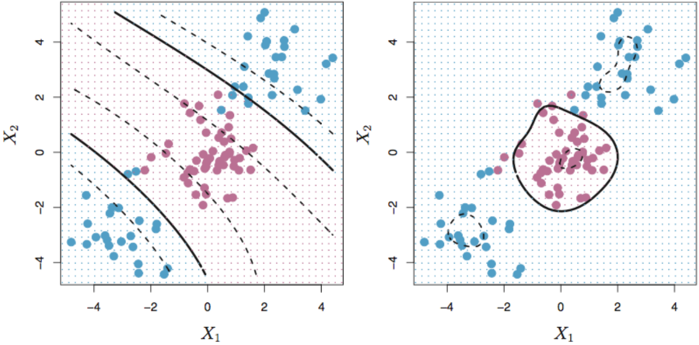
$$s.t. \quad y_i(w^T x_i + b) \geq 1 - \epsilon_i,$$

$$\epsilon_i \geq 0$$

$$i = 1, \ldots, n$$

Support Vector Machine

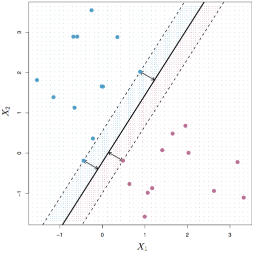
An extension of Support Vector Classifiers that enlarges the feature space using kernels to create a non-linear decision boundary



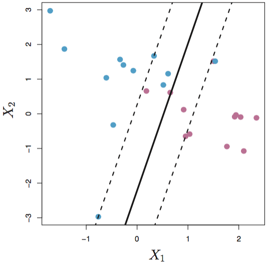
Different Dual Optimization problems depending on choice of kernel, notably only depending on the inner products of observations

Support Vector Machine

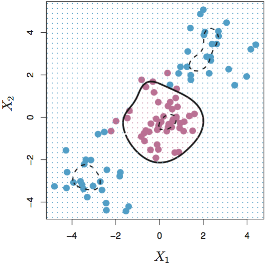
Maximal Margin Classifiers, Support Vector Classifiers, and Support Vector Machines are all considered Support Vector Machines



Linear Kernel, $\epsilon_i = 0$



Linear Kernel, $\epsilon_i > 0$



Radial Kernel

Naive Bayes Classifier

Given a vector $\mathbf{x} = (x_1, ..., x_n)^T$, We assign the probability

$$P(C_k|x_1, ...x_n)$$

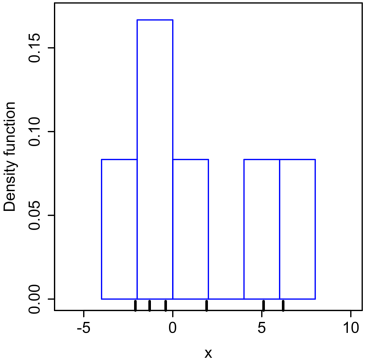
to the event that the observation x_i belongs to the class C_k. We assume each feature is conditionally independent of every other feature given the class variable.

Using Bayes' theorem, the Naive Bayes classifier is the following function that assigns the observation to the class:

$$\hat{y} = \underset{k \in \{1,...,K\}}{argmax} P(C_k) \prod_{i=1}^{m} P(x_i|C_k)$$

Kernel Density Estimation

Next, we want to know how to calculate the conditional probability $P(x_i|C_k)$ in a non-parametric way



Using histograms, we can estimate the probability as

$$\hat{f}(x_0) = \frac{\#x_i \in \mathcal{N}(x_0)}{nh}$$

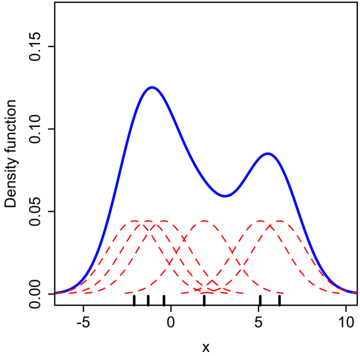
where $h > 0$ is a parameter called the bandwidth

Kernel Density Estimation

Using kernels we can obtain a *smooth* estimate for the pdf

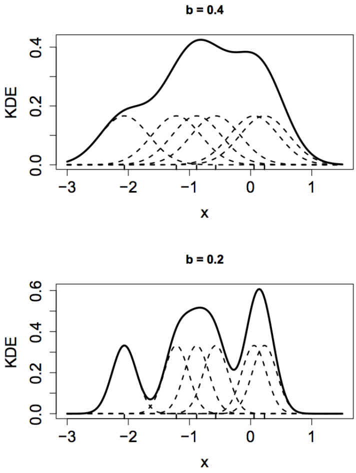
$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where $h > 0$ is the bandwidth, and $K(u)$ is the kernel function



Bias-Variance Tradeoff

The choice of bandwidth h is important because of the bias-variance tradeoff

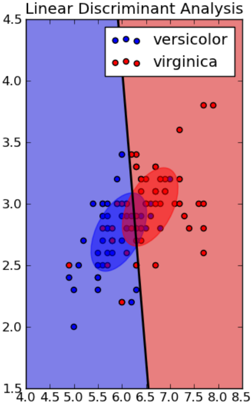


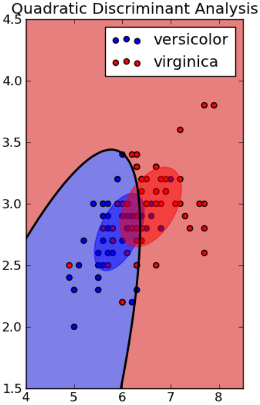
Overall of Our Empirical Studies

- Six Individual Empirical Studies:
- Heart Disease Data Analysis — Andy Wu
- Text Classification(BBC News Data Set...) — Shiyuan Li
- Categorical Predictors(Connect-4 Data Set...) — Xi Cheng
- Sentiment Analysis(IMDB Reviews Data Set...) — Zimeng Wang
- SVM for Unbalanced Data — Jing Peng
- Connection between SVM, LDA and QDA — Heng Xu

Connection between SVM, LDA and QDA

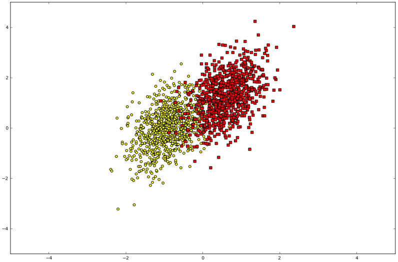
What is LDA and QDA:





Connection between SVM, LDA and QDA

- When we want use LDA and QDA?



- LDA:
  Assuming each class has the same variance - covariance matrices.
  Straight Line.
- QDA:
  Assuming each class has different variance - covariance matrices.
  Quadratic Curve.

Covariance Adjusted SVM

Linear SVM with Soft-Margin:

$$\begin{aligned} \underset{w,b,\varepsilon_i}{\text{minimize}} \quad & \frac{1}{2}w^T w + C\sum_{i=1}^{n}\varepsilon_i \\ \text{s.t.} \quad & y_i(w^T x_i + b) \geq 1 - \varepsilon_i, \text{ and } \varepsilon_i \geq 0, i = 1, \ldots, n, \end{aligned} \tag{1}$$

Dual form of Kernel SVM:

$$\begin{aligned} \max_{\alpha_i \geq 0} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K(x_i, x_j) \\ s.t. \quad & \sum_{i=1}^{n} y_i\alpha_i = 0 \text{ and } 0 \leq \alpha_i \leq C, \text{for } i = 1, ..., n \end{aligned} \tag{2}$$

Covariance Adjusted SVM

We here use S to denote the pooled covariance matrix and we want to add variance-covariance into our consideration:

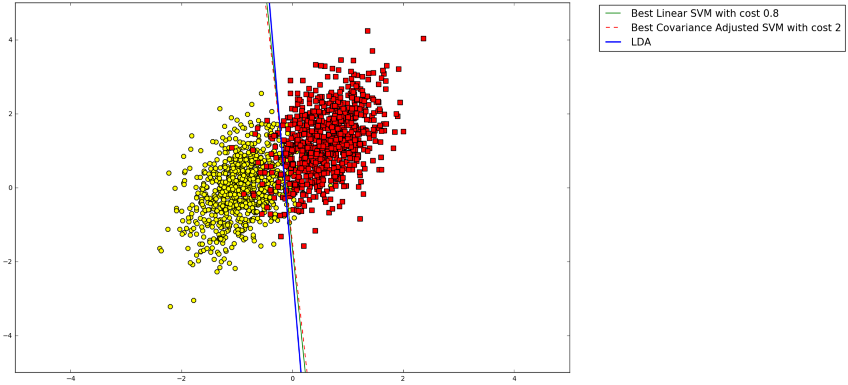
$$\underset{w,b,\varepsilon_i}{\text{minimize}} \quad \frac{1}{2}w^T S w + C\sum_{i=1}^{n}\varepsilon_i$$
$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \varepsilon_i, \text{ and } \varepsilon_i \geq 0, i = 1, \ldots, n,$$

We can verify that this model is equivalent to multiply the inverse of the square root of pooled covariance matrix to our data, and then apply SVM to the new data:

$$\underset{w,b,\varepsilon_i}{\text{minimize}} \quad \frac{1}{2}\mathbf{w}^T (S^{\frac{1}{2}})^T S^{\frac{1}{2}}\mathbf{w} + C\sum_{i=1}^{n}\varepsilon_i,$$
$$\text{s.t.} \quad y_i(\mathbf{w}^T S^{\frac{1}{2}} S^{-\frac{1}{2}} x_i + b) \geq 1, \text{for i} = 1,\ldots,\text{n},$$

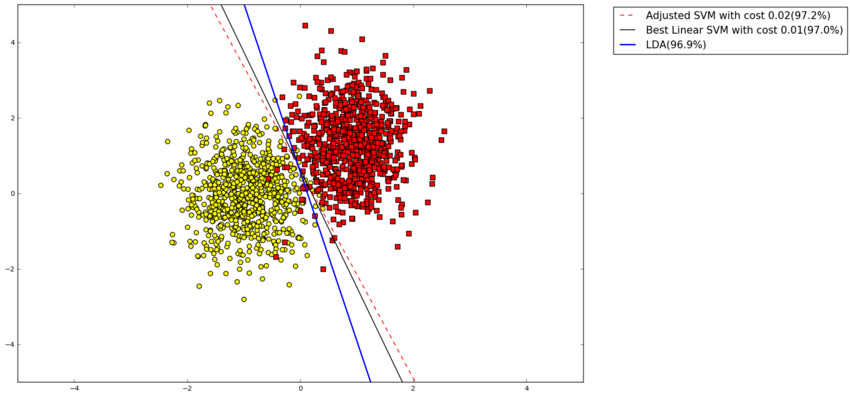
Connection between SVM, LDA and QDA

- Case 1: 2 dimension, Same Variance-Covariance matrix and merged heavily



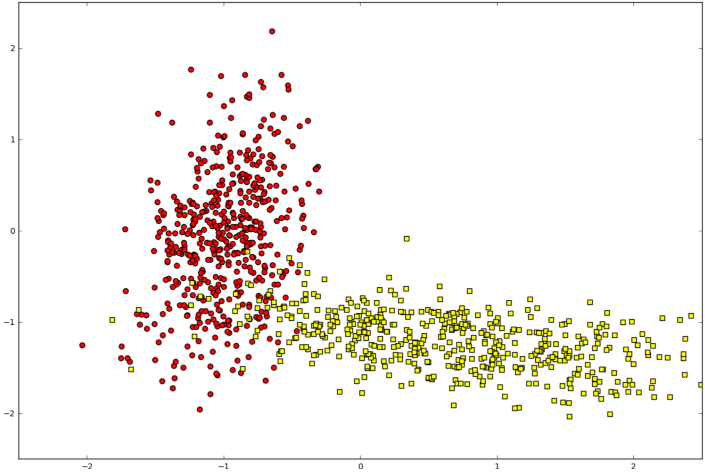
Connection between SVM, LDA and QDA

- Case 2: 2 dimension, Same Variance-Covariance matrix but not merged heavily



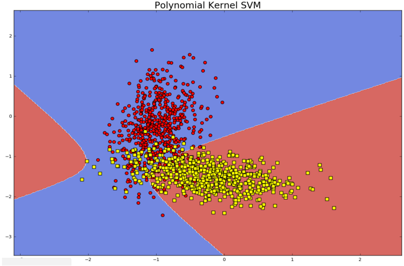
Connection between SVM, LDA and QDA

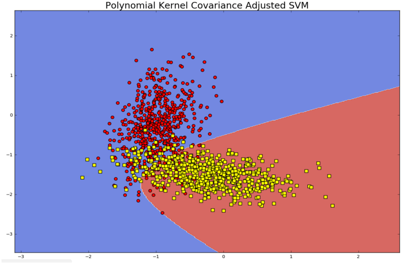
- 2 dimension, different variance-covariance matrix(using SVM with polynomial kernel of degree 2 and QDA)

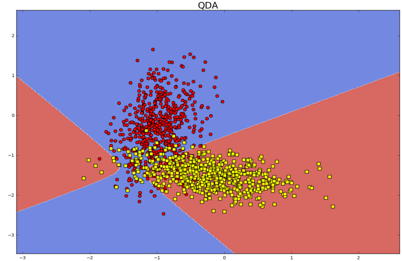


Connection between SVM, LDA and QDA

- Case 3: If two classes are mixed with each other

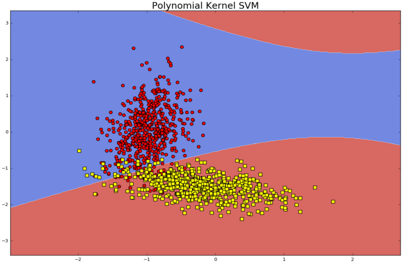


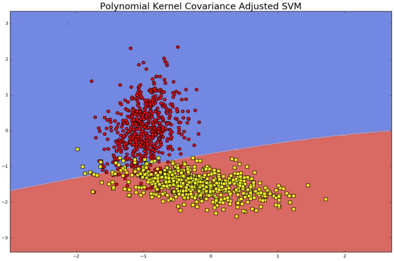


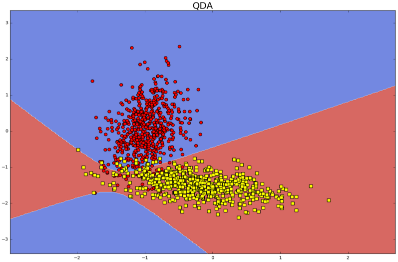


Connection between SVM, LDA and QDA

- Case 5: If two classes are not mixed such heavily

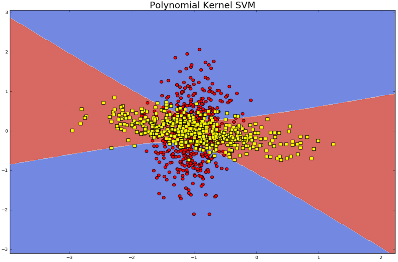


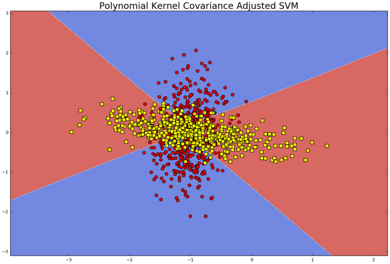


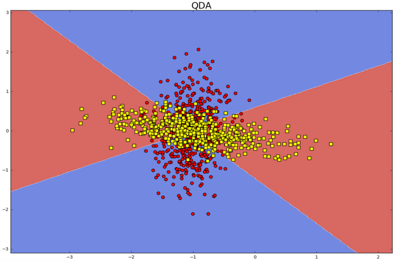


Connection between SVM, LDA and QDA

- Case 4: If mixed heavily, even in some extreme cases



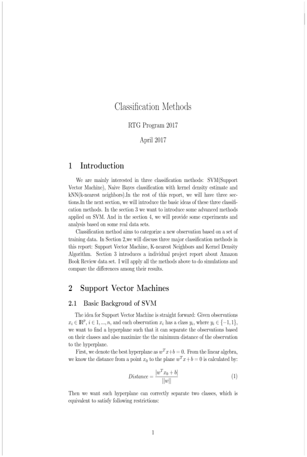




Connection between SVM, LDA and QDA

- Opinions: If two classes twisted with each other a lot, linear SVM and LDA(Polynomial Kernel SVM and QDA) will construct extreme similar classifiers.

- We are writing all of our thoughts and work in an INTERESTING report here!



Classification Methods

RTG Program 2017

April 2017

1 Introduction

We are mainly interested in three classification methods: SVM(Support Vector Machine), Naive Bayes classification with kernel density estimate and kNN(k-nearest neighbors).In the rest of this report, we will have three sections.In the next section, we will introduce the basic ideas of these three classification methods. In the section 3 we want to introduce some advanced methods applied on SVM. And in the section 4, we will provide some experiments and analysis based on some real data sets.

Classification method aims to categorize a new observation based on a set of training data. In Section 2,we will discuss three major classification methods in this report: Support Vector Machine, K-nearest Neighbors and Kernel Density Algorithm. Section 3 introduces a individual project report about Amazon Book Review data set. I will apply all the methods above to do simulations and compare the differences among their results.

2 Support Vector Machines

2.1 Basic Backgroud of SVM

The idea for Support Vector Machine is straight forward: Given observations $x_i \in \mathbb{R}^p$, $i \in 1, ..., n$, and each observation x_i has a class y_i, where $y_i \in \{-1, 1\}$, we want to find a hyperplane such that it can separate the observations based on their classes and also maximize the the minimum distance of the observation to the hyperplane.

First, we denote the best hyperplane as $w^T x + b = 0$. From the linear algebra, we know the distance from a point x_0 to the plane $w^T x + b = 0$ is calculated by:

$$Distance = \frac{|w^T x_0 + b|}{||w||} \quad (1)$$

Then we want such hyperplane can correctly separate two classes, which is equivalent to satisfy following restrictions:

1