Binary Classification with Missing Data

Huafeng Fan

May 6, 2019

Reference

- Bhattacharyya, Chiranjib, Pannagadatta K. Shivaswamy, and Alex J. Smola. "A second order cone programming formulation for classifying missing data." Advances in neural information processing systems. 2005.

Overview

Problem Statement and Approach

- In class, we looked at binary classification using SVM in the case where the training data and labels were known.
- In practice, datasets are messy. Even on Kaggle, there are many datasets with incomplete and noisy data. We would like to still achieve optimal classification results with noisy data.
- The approach to solve this problem is to assume our data takes on a Gaussian distribution, and derive a convex optimization problem for classification. This problem will turn out to be a SOCP.

Support Vector Machine

- Given training data $\{x_i, y_i\}_{i=1}^m$ where $x_i \in \mathbb{R}^n$, $y_i \in \{-1, 1\}$, we have two forms of binary classifiers from class:
- LP Heuristic (with an added norm constraint)

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m u_i \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 - u_i (\forall i = 1, \ldots, m) \\ & u_i \geq 0 \qquad\qquad\qquad\quad (i = 1, \ldots, m) \\ & \|w\|_2 \leq W \end{array} \tag{1}$$

- Standard Support Vector Machine classifier

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\|w\|_2^2 + C \sum_{i=1}^m u_i \\ \text{subject to} & y_i(w^T x_i + b) \geq 1 - u_i (\forall i = 1, \ldots, m) \\ & u_i \geq 0 \qquad\qquad\qquad\quad (\forall i = 1, \ldots, m) \end{array} \tag{2}$$

- Continue with form (1) to derive the problem for when x_i isn't exactly known.

Dealing with unknown x_i

- In the case where we don't know x_i, we can assume $x_i \sim P_i$, some probability distribution for all i.
- Now our training data is $\{x_i, y_i\}_{i=1}^m$ where $x_i \sim P_i$, $y_i \in \{-1, 1\}$.
- It makes sense now to take probabilities into account, giving us the problem

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^m u_i \\ \text{subject to} & \Pr\{y_i(w^T x_i + b) \geq 1 - u_i\} \geq k_i (\forall i = 1, \ldots, m) \\ & u_i \geq 0 \qquad\qquad\qquad\qquad\qquad (i = 1, \ldots, m) \\ & \|w\|_2 \leq W \end{array} \tag{3}$$

  where k_i is user defined.
- To simplify, we need to know the distributions of x_i.

Dealing with unknown $x_i \sim N(\bar{x}_i, \Sigma_i)$

- Suppose $x_i \sim N(\bar{x}_i, \Sigma_i)$. Then, $z_i = y_i(w^T x_i + b) \sim N(\bar{z}_i, \sigma_{z_i}^2)$. where $\bar{z}_i = y_i(w^T \bar{x}_i + b)$ and $\sigma_{z_i}^2 = w^T \Sigma_i w$, we have that

$$\begin{aligned}
\Pr\{y_i(w^T x_i + b) \geq 1 - u_i\} &= \Pr\{z_i \geq 1 - u_i\} \\
&= \Pr\{\frac{z_i - \bar{z}_i}{\sigma_{z_i}} \geq \frac{1 - u_i - \bar{z}_i}{\sigma_{z_i}}\} \\
&= \phi(\frac{\bar{z}_i + u_i - 1}{\sigma_{z_i}}) \geq k_i \\
&\implies \bar{z}_i \geq \phi^{-1}(k_i)\sigma_{z_i} - u_i + 1 \\
&\implies y_i(w^T \bar{x}_i + b) \geq 1 - u_i + \gamma_i \sigma_{z_i} \\
&\implies y_i(w^T \bar{x}_i + b) \geq 1 - u_i + \gamma_i \sqrt{w^T \Sigma_i w}
\end{aligned}$$

where $\phi(u) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp(-\frac{s^2}{2}) ds$ and $\gamma_i := \phi^{-1}(k_i)$

Deriving the SOCP Problem

- Putting it all together so far, we have the optimization problem:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m} u_i \\ \text{subject to} & y_i(w^T\bar{x}_i + b) \geq 1 - u_i + \gamma_i\sqrt{w^T\Sigma_i w} \quad (\forall i = 1, \ldots, m) \\ & u_i \geq 0 \qquad (i = 1, \ldots, m) \\ & ||w||_2 \leq W \end{array} \tag{4}$$

- Noting that all covariance matrices are positive semi-definite, we get the problem:

$$\begin{array}{ll} \text{minimize} & \sum_{i=1}^{m} u_i \\ \text{subject to} & y_i(w^T\bar{x}_i + b) \geq 1 - u_i + \gamma_i||\Sigma_i^{1/2}w||_2 \quad (\forall i = 1, \ldots, m) \\ & u_i \geq 0 \qquad (i = 1, \ldots, m) \\ & ||w||_2 \leq W \end{array} \tag{5}$$

- There are three cases for the value of $\gamma_i = \phi^{-1}(k_i)$
  - $\gamma_i = 0$ or $k_i = 0.5$: Original SVM problem.
  - $\gamma_i < 0$ or $k_i < 0.5$: Hard optimization problem
  - $\gamma_i > 0$ or $k_i > 0.5$: SOCP

SOCP Problem with both Known and Unknown Data

- Let m_a be the number of datapoints for which the values are available.
- Let m_m be the number of datapoints containing missing values.
- The optimization problem in this case is:

$$
\begin{array}{lll}
\text{minimize} & \sum_{i=1}^{m} u_i & \\
\text{subject to} & y_i(w^T x_i + b) \geq 1 - u_i & (i = 1, \ldots, m_a) \\
 & y_i(w^T \bar{x}_i + b) \geq 1 - u_i + \gamma_i ||\Sigma_i^{1/2} w||_2 & (i = m_a + 1, \ldots, m_a + \\
 & u_i \geq 0 & (i = 1, \ldots) \\
 & ||w||_2 \leq W &
\end{array}
\tag{6}
$$

- We can estimate $\bar{x}_i$ and Σ_i using the Expectation Minimization (EM) algorithm from our known data assuming that x follows a jointly normal distribution with mean μ and covariance Σ.

Deriving the Dual Problem 1

- The Lagrangian of the program on the previous slide is:

$$L(w, b, u, \alpha, \beta, \lambda, \theta) = 1^T u + \sum_{i=1}^{m_a} \alpha_i [1 - u_i - y_i(w^T x_i + b)] +$$
$$\sum_{i=m_a+1}^{m_a+m_m} \beta_i [1 - u_i - y_i(w^T \bar{x}_i + b)] +$$
$$\sum_{i=m_a+1}^{m_a+m_m} \beta_i \gamma_i ||\Sigma_i^{1/2} w||_2 -$$
$$\sum_{i=1}^{m_a+m_m} \lambda_i u_i + \theta(||w||_2 - W)$$

where $\alpha, \beta, \lambda, \theta$ are the Lagrange multipliers.

Deriving the Dual Problem 2

- Deriving the Lagrange Dual Function $g(\alpha, \beta, \lambda, \theta) = \inf_{w,b,u} L(w, b, u, \alpha, \beta, \lambda, \theta)$ as in class, we get:

$$g(\alpha, \beta, \lambda, \theta) = \begin{cases} -W\theta & \text{if } B \leq \theta + \sum_{i=m_a+1}^{m_a+m_m} \beta_i ||\Sigma_i^{1/2}||_2 \\ -\infty & \text{otherwise} \end{cases}$$

  where

$$B = || - \sum_{i=1}^{m_a} \alpha_i y_i x_i - \sum_{i=m_a+1}^{m_a+m_m} \beta_i y_i \bar{x}_i ||_2$$

- This is similar to the dual problem of the normal SVM problem.

Deriving the Dual Problem 3

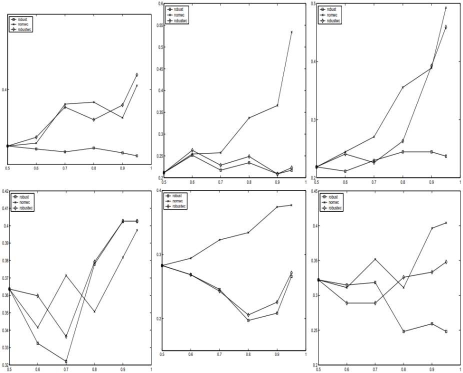
- Finally, we get the dual problem:

$$\begin{array}{ll} \text{maximize} & ||-\sum_{i=1}^{m_a}\alpha_i y_i x_i - \sum_{i=m_a+1}^{m_a+m_m}\beta_i y_i \bar{x}_i||_2 \\ \text{subject to} & 1^T\alpha = 1^T\beta = \frac{1}{2} \\ & \lambda \succeq 0 \\ & \theta \succeq 0 \end{array} \tag{7}$$

- We can do sensitivity analysis on how good the classifier is wrt how inseparable and how uncertain we are about the data.

Experimental Results

- From Bhattacharyya, Pannagadatta, Smola's experiments on the public datasets Prima, Heart and Ionosphere:



- Results were mostly verified by hand, will try on more datasets and more graphs will be procured for the final report.