Semi-supervised Principal Components Regression

Charles Zheng

April 22, 2015

Abstract

Semi-supervised learning refers to the problem of learning a rule for predicting labels y from features x, given training data which includes both labeled examples $(x_1, y_1), \ldots, (x_\ell, y_\ell)$ and unlabeled examples $x_{\ell+1}, \ldots, x_n$. A basic question of this field is to characterize the conditions under which the unlabeled examples can be used to improve generalization error. We introduce a latent variable model for which semi-supervised principal components regression is shown to outperform supervised ridge regression.

Keywords: Semi-supervised, ridge regression, principal components analysis, latent variables

1 Introduction

1.1 Ridge Regression

Ridge regression is a linear method for predicting a real-valued label $y \in \mathbb{R}$ from a vector of real-valued features $x \in \mathbb{R}^p$. It is produces a regularized least squares estimate for a coefficient vector $\hat{\beta}_\lambda$ and intercept term $\hat{c}_\lambda$, where λ is a regularization parameter. This coefficient vector $\hat{\beta}_\lambda$ is used to predict the label y^* for an unlabeled example with features x^* by the rule

$$y^* = \hat{\beta}_\lambda^T x^* + \hat{c}_\lambda$$

Given training data with labels $Y = (y_1, \ldots, y_n)$ and features $X = (x_1^T, \ldots, x_n^T)$, the ridge regression

1.2 Principal Components Regression

2 Theory

2.1 Model

We specify a generative model for data matrices $X = (x_1^T, \ldots, x_n^T)$ and $Y = (y_1, \ldots, y_n)$, which together represent n labeled examples.

Let $Z = (z_1^T, \ldots, z_n^T)$ be an $n \times r$ matrix of latent variables, where each entry z_{ij} is iid standard normal. The latent variables are related to X and Y in the following way. Let

$$X = Z\alpha + 1_n C^T + E$$

where α is a $r \times p$ coefficient matrix with unit-norm columns, C is a fixed $p \times 1$ intercept matrix, and E is a $n \times p$ random matrix of error terms which are iif $N(0, \sigma_\epsilon^2)$. Similarly, let

$$Y = Z\gamma + c + \epsilon$$

where γ is a $r \times 1$ coefficient vector, c is an intercept term, and ϵ is a $n \times 1$ vector with entries iid $N(0, \sigma_\epsilon^2)$.

The semi-supervised learning problem can be posed as follows. Let $\ell < n$ be the number of labeled examples, and let Y_ℓ be the first ℓ rows of Y. Supposing that only X and Y_ℓ are observed, and α, γ, and σ_ϵ^2 are unknown parameters, the problem is to predict $\hat{Y}$ for all n examples. The squared-error loss is defined as

$$||\hat{Y} - Y||^2$$

and the goal is to choose a prediction method which minimizes the *risk*, or expected squared-error loss.

Of course, one can always set the first ℓ entries of $\hat{Y}$ to be equal to Y_ℓ, which guarantees zero error on those examples. Thus the challenge is to make a prediction for the unobserved entries of Y.

Throughout the paper we make the simplifying assumption that r, the number of latent variables, is known; however, the problem of testing the number of principal components r is well-studied in the statistics literature, and our analysis could be extended to incorporate the case of unknown r. Noting that the problems of estimating the unknown intercept terms C and c as well as the marginal variances of X and Y are also well-understood, we lose little theoretical power and gain much clarity by making additional assumptions that $C = 0$, $c = 0$, and that all the columns of α and γ are unit-norm. Note that as a result, X and Y have zero marginal mean and equal marginal variances of $1 + \sigma_\epsilon^2$, justifying the omission of the normalization step normally employed in ridge regression.

3 References

- Zhu, X. “Semi-supervised learning literature survey.” 2005.
- Niyogi, P. “Manifold Regularization and semi-supervised learning: Some theoretical analysis.” 2008.