10-701/15-781, Machine Learning: Homework 3

Eric Xing, Tom Mitchell, Aarti Singh
Carnegie Mellon University
Updated on February 7, 2010

1 Linear regression, and bias-variance trade-off[20pt, Ni Lao]

In linear regression, we are given training data of the form, $\mathcal{D} = (\mathbf{X}, \mathrm{y}) = \{(\mathrm{x}_i, y_i)\}, i = 1, 2, ..., n$, where $\mathrm{x}_i \in \mathcal{R}^{1\times p}$, i.e. $\mathrm{x}_i = (x_{i,1}, \cdots, x_{i,p})^{\mathrm{T}}$, $y_i \in \mathcal{R}$, $\mathbf{X} \in \mathcal{R}^{n\times p}$, where n is number of samples, p is number of features. Each row i of $\mathbf{X}$ is ${\mathrm{x}_\mathrm{i}}^{\mathrm{T}}$, and $\mathrm{y} = (y_1, \cdots, y_n)^{\mathrm{T}}$.

Least square regression seeks to find β that minimizes the square-error, i.e.:

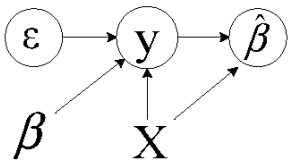
$$\mathrm{J}_1(\beta) = \sum_i (y_i - {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta)^2 = (\mathbf{X}\beta - \mathrm{y})^{\mathrm{T}}(\mathbf{X}\beta - \mathrm{y}).$$

It has an unique solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathrm{y}. \tag{1}$$

where $\hat{\beta}$ is called an estimator of β. An "estimator" is a statistic of your data (i.e. a function of your data) which is intended to approximate a parameter of the underlying distribution. There is a research field called "estimation theory", which deals with constructing estimators that have nice properties, like converging to the correct parameter given enough data, and giving confidence intervals. In this problem we will explore how regularization affects the bias and variance of the least square regression model.

Let's assume that $p < n$, and $\mathbf{X}^T\mathbf{X}$ is invertible. Also assume that our data is generated from a true model of the form: $y_i = {\mathrm{x}_\mathrm{i}}^{\mathrm{T}}\beta + \epsilon_i$ (or in matrix form $\mathrm{y} = \mathbf{X}\beta + \epsilon$), where $\epsilon_1, ..., \epsilon_n$ are IID and sampled from a Gaussian with 0 mean and constant standard deviation, that is $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ (or $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$). Relations among these quantities can be summarized by the figure below, where quantities with circles are random variables. Here we assume that $\mathbf{X}$ is "deterministic" and fixed.



1.1 Least square regression [4 pt]

Show that $\hat{\beta}$ has Gaussian distribution and write down its mean μ and covariance matrix Σ. You will see that least square regression is unbiased $E[\hat{\beta}] = \beta$

Hint: if $\epsilon \sim \mathcal{N}(0, I)$ then $A\epsilon \sim \mathcal{N}(0, AA^T)$, where A is any matrix.

Hint: notation is simpler, if you do a Singular Value Decomposition (SVD) to $\mathbf{X}$ first.

1.2 Ridge regression [4pt]

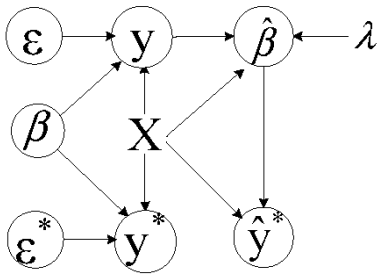
The solution to ridge regression is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathrm{y}. \tag{2}$$

Show that $\hat{\beta}$ has Gaussian distribution and write down its mean μ and covariance matrix Σ. You will see that ridge regression is biased $E[\hat{\beta}] \neq \beta$.

1.3 The bias variance trade-off [4 pt]

Now let's be real Bayesians, and believe that the true parameter β itself is a random variable. Let's assume that $\beta \sim \mathcal{N}(0, \alpha^2 I)$. Apart from our training data $\mathcal{D} = (\mathbf{X}, \mathrm{y})$, we also generate a set of testing data $\mathcal{D}^* = (\mathbf{X}, \mathrm{y}^*)$. It has exactly the same x values $\mathbf{X}$ as training data, but the y values are regenerated independently. Again we can decompose them as $\mathrm{y}^* = \mathbf{X}\beta + \epsilon^*$. Relations among these quantities can be summarized by the figure below.



Now we want see how should we choose the regularization parameter λ so that the risk of ridge regression on test data $\mathcal{D}^*$ is minimized.

As a first step, let's express $\mathrm{e}(\lambda) = \hat{\mathrm{y}}^* - \mathrm{y}^*$, the test errors of ridge regression on $\mathcal{D}^*$, in terms of β, ϵ, and ϵ^*. You will see three terms: one grows as λ grows corresponding to the bias of our estimation, one diminishes as λ grows corresponding to the variance of our estimation, and one that is independent of λ corresponding to the irreducible error.

1.4 [4 pt]

Express the risk of ridge regression $R(\lambda) = E[\mathrm{e}(\lambda)^T \mathrm{e}(\lambda)] = \sum_i E[e_i^2(\lambda)]$ in terms of α, σ, and λ.

Hint: if $a \sim \mathcal{N}(0, ASA^T)$, where A is unitary, S is diagonal, then $E[a^T a] = \sum_i S_{i,i} = tr(S)$.

Hint: if a and b are independent random vectors with zero means, then $E[(a+b)^T(a+b)] = E[a^T a] + E[b^T b]$

1.5 [4 pt]

Find the optimal regularization parameter λ, that minimizes the risk of ridge regression on test data $\mathcal{D}^*$. You will see that the larger the magnitude of data noise ϵ (controlled by σ) and the smaller the magnitude of the true parameter β (controlled by α) the larger regularization is needed to achieve the minimum risk.

Hint: for simplicity, you can assume that the number of features $p = 1$. The result should still hold for $p > 1$.

Hint: the result is an expression of σ and α.