CS 189 Introduction to Machine Learning
Fall 2017

gs1

1 Frobenius Norm, Trace, SVD

The Frobenius norm of a matrix is defined as the square root of the sum of the absolute squares of its elements:

$$\|M\|_F = \sqrt{\sum_{i,j} |m_{i,j}|^2}$$

In class we have used the following property to simplify our calculations of the Frobenius norm,

$$\|M\|_F = \sqrt{\mathrm{Tr}(MM^T)}$$

(a) Prove that the above equation is true.

(b) Now show that the Frobenius norm of a square matrix is equal to the square root of the sum of the singular values:

$$\|M\|_F = \sqrt{\sum_i \sigma_i^2}$$

(c) In HW 4 we used the Eckart-Young-Mirsky thereom to find the closest lower-rank approximation of a degenerate matrix in the Frobenius Norm. Show that the following is true:

Given a matrix $M \in R^{mxn}$ with singular values $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_n$

$$\min_{\tilde{M}_k} \|M - \tilde{M}_k\|_F^2 = \sum_{i=k+1}^{m} \sigma_i^2$$

$Hint$: First show that i) There is a rank k matrix $\tilde{M}_k$ such that

$$\|M - \tilde{M}_k\|_F^2 = \sum_{i=k+1}^{n} \sigma_i^2$$

then show that ii) $\tilde{M}_k$ yields the minimal solution to the optimization problem.

2 Bias/Variance for K-Nearest Neighbors Regression

Suppose we have n training points x_i with labels y_i. We want to model a regression problem with k-nearest neighbors regression. K-nearest neighbors works as follows: for a particular data point z, the k-nearest neighbors regression algorithm finds the closest k points to z in our n training points and predicts the value label for z by averaging the labels of the closest k points. More formally, we model our hypothesis $h(z)$ as

$$h(z) = \frac{1}{k}\sum_{i=1}^{n} N(x_i, z, k)$$

where the function N is defined as

$$N(x_i, z, k) = \begin{cases} y_i & \text{if } x_i \text{ is one of the } k \text{ closest points to } z \\ 0 & o.w. \end{cases}$$

Suppose also we assume our labels $y_i = f(x_i) + \varepsilon$, where ε is the noise that comes from $\mathcal{N}(0, \sigma^2)$ and f is the true function.

(a) Derive the bias^2 of our model for given x_i, y_i pairs. Remember that the bias is simply $(\mathbb{E}(h(z)) - f(z))^2$.

(b) How well does k-nearest neighbors behave as $k \longrightarrow \infty$? How about when $k = 1$? Comment.

(c) Derive the variance of our model, which is defined as the $Var(h(z))$.

(d) What happens to the variance when $k \longrightarrow \infty$? How about when $k = 1$?

3 MLE, MAP, and Lasso

Assume a set of points $x_1, \ldots, x_n \in \mathbb{R}^d$, an unknown parameter vector $\theta^* \in \mathbb{R}^d$, and observations $y_1, \ldots, y_n \in \mathbb{R}$ generated by

$$y_i = x_i^\top \theta^* + \varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ for some arbitrary σ^2. Note that this can equivalently be written as

$$y_i \sim \mathcal{N}(x_i^\top \theta^*, \sigma^2)$$

(a) Show that performing maximum likelihood estimation under these modeling assumptions is equivalent to solving the unconstrained least squares problem. That is, show that you can formulate the optimization problem as

$$\hat{\theta} = \arg\min_{\theta} \alpha \|X\theta - Y\|_2^2 \tag{1}$$

for $\alpha > 0$, $X \in \mathbb{R}^{n \times d}$, $Y \in \mathbb{R}^n$.

(b) Now assume that θ_i^* is drawn from a distribution with probability density function $p(\theta_i^*) \propto e^{-|\theta_i^*|/t}$ where $t > 0$ is a constant. Show that performing maximum a posteriori estimation is equivalent to solving the l-1 regularized least squares problem. That is, show that you can formulate the optimization problem as

$$\hat{\theta} = \arg\min_{\theta} \alpha \|X\theta - Y\|_2^2 + \beta \|\theta\|_1 \tag{2}$$

for $\alpha > 0$, $\beta > 0$, $X \in \mathbb{R}^{n \times d}$, $Y \in \mathbb{R}^n$.

(c) Consider the following l-2 regularized regression problem:

$$\hat{\theta} = \arg\min_{\theta} \|X\theta - Y\|_2^2 + \lambda \|\theta\|_2^2 \tag{3}$$

Solve for $\hat{\theta}$ and show that it is a biased estimator.

(d) Consider the optimization problem below that combines l-1 and l-2 regularization with $\gamma \in [0, 1]$:

$$\hat{\theta} = \arg\min_{\theta} \|X\theta - Y\|_2^2 + \lambda \left[\gamma \|\theta\|_2^2 + (1-\gamma)\|\theta\|_1\right] \tag{4}$$

Show that it can be rewritten as an l-1 regularized problem with augmented versions of X and Y.

Hint: You can modify X to be a specific block matrix.

4 GLS and the Gauss-Markov Theorem

Suppose we are in the GLS setting where we have a model $Y = Xw + N$ where $N \sim \mathcal{N}(0, \Sigma)$ for some PSD covariance matrix Σ (that is, the error terms could be correlated). Recall that the GLS estimate is $\hat{w}_{GLS} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Y$ and coincides with the MLE when N is Gaussian. In this problem we will show that the GLS estimator is a "best linear unbiased estimator" of w in that it yields the lowest mean squared error $E(\|\hat{w} - w\|_2^2)$ out of all unbiased estimators $\hat{w}$ of w that are linear in y.

(a) Compute $E(\hat{w}_{GLS})$ and $Cov(\hat{w}_{GLS})$. What is the distribution of $\hat{w}$?

(b) Show that $MSE(\hat{w}) = E(\|w - \hat{w}\|_2^2)$ can be decomposed into the sum of the squared norm of the bias, $\|w - E(\hat{w})\|_2^2$, and the trace of the covariance matrix $\mathrm{Tr}(Cov(\hat{w}))$. Conclude that for unbiased estimators $\hat{w}$ of w, $MSE(\hat{w}) = \mathrm{Tr}(Cov(\hat{w}))$.

(c) In this part of the problem we will prove a version of the Gauss-Markov Theorem for GLS, which states that if $\hat{w}$ is an unbiased estimator of w that is linear in y (that is, $\hat{w} = Cy$ for some C), then $Cov(\hat{w}) - Cov(\hat{w}_{GLS})$ is positive semi-definite.

  (a) Set $M = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}$ so that $\hat{w}_{GLS} = MY$. If $\hat{w} = (M + D)Y$ where $D \neq 0$ (because if $D = 0$, $\hat{w} = \hat{w}_{GLS}$), show that a necessary and sufficient condition for $\hat{w}$ to be unbiased for every w is the condition $DX = 0$ (hint: take $E(\hat{w})$ and express it as β plus another term).

  (b) Show that $Cov(\hat{w}_{GLS}) - Cov(\hat{w})$ is PSD for every such $\hat{w}$ satisfying the conditions for the Gauss-Markov Theorem (hint: take $Cov(\hat{w})$ and express it as $Cov(\hat{w}_{GLS})$ plus another term using the condition found in part (a) - then show that term is PSD).

  (c) Does the Gauss-Markov theorem apply when the errors N do not follow a normal distribution?

(d) Conclude that the GLS estimator minimizes the MSE over all unbiased estimators that are linear in y. In particular, if the covariance matrix of the errors is not a multiple of the identity, GLS does at least as well as OLS.