Homework #4

CS 260: Machine Learning Algorithms
Prof. Ameet Talwalkar
Due: 10/29/15, 8am
Please abide by the Academic Integrity Policy

1 Linear Regression with Heterogenous Noise

In the standard linear regression model, we consider the model that the observed response variable y is the prediction perturbed by noise, namely

$$y = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon$$

where ε is a Gaussian random variable with mean 0 and variance σ^2. Notably, we are assuming that for all observations in the training data, the corresponding noises are identically and independently distributed. In other words, for the n-th observation $\mathbf{x}_n$, the observed response is

$$y_n = \mathbf{x}_n^\top \boldsymbol{\beta} + \varepsilon_n$$

where $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$.

This assumption is not applicable in some cases. For example, in the example of predicting the sale prices of houses, the variances for larger houses (e.g., houses with larger $\mathbf{x}_n$ which is the square footage) tend to be bigger, as the sale prices for larger houses seem to be more variable.

In this case, we can model the data in the following way:

$$y_n = \mathbf{x}_n^\top \boldsymbol{\beta} + \varepsilon_n$$

where ε_n are independently distributed but **do not have to be identically distributed**. In particular, each one could have a different variance, namely, $\varepsilon_n \sim \mathcal{N}(0, \sigma_n^2)$.

(a) Suppose our training dataset contains $\{(\mathbf{x}_n, y_n), n = 1, 2, \ldots, N\}$ such observations. Write down the log-likelihood function of the data. This function should be a function of the data as well as $\boldsymbol{\beta}$ and all σ_n.

(b) Derive the maximum likelihood estimate of $\boldsymbol{\beta}$, and express it in terms of the data as well as all the σ_n. You should assume σ_n is known to you — you do not need to estimate them from the data.

2 Linear Regression with Smooth Coefficients

Consider a dataset with n data points $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathbb{R}^{p \times 1}$, drawn from the following linear model:

$$y = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon,$$

where ε is a Gaussian noise. Suppose the features $x_{i1}, \ldots, x_{ip}$ for all $i = 1, \ldots, n$ have a natural ordering. Several examples have this ordering property; for example in the study of the impact of proteins on certain types of cancer, the proteins are ordered sequentially on a line. Intuitively, we can encode the natural ordering information by introducing a condition that requires the difference $(\beta_i - \beta_{i+1})^2$ cannot be large, for $i = 1, \ldots, p-1$.

(a) State the condition as a regularizer. Write the new optimization problem for finding $\boldsymbol{\beta}$ by combining both this regularization and L_2 regularization. (10 points)

(b) Find the optimal $\boldsymbol{\beta}$ by solving the problem in part (a). (5 points)

3 Linearly Constrained Linear Regression

Consider a dataset with n data points $(\mathbf{x}_i, y_i)$, $\mathbf{x}_i \in \mathbb{R}^{p\times 1}$, drawn from the following linear model:

$$y = \mathbf{x}^\top \boldsymbol{\beta} + \varepsilon,$$

where ε is Gaussian noise. Suppose we have additional information about $\boldsymbol{\beta}$ that requires $A\boldsymbol{\beta} = \mathbf{b}$ where $A \in \mathbb{R}^{q\times p}$ and $\mathbf{b} \in \mathbb{R}^{q\times 1}$. Suppose the constraint $A\boldsymbol{\beta} = \mathbf{b}$ has a non-empty set of solutions; thus the optimization has feasible solutions. Find the maximum likelihood estimation of $\boldsymbol{\beta}$ under this constraint.

4 Online Learning

The perceptron algorithm often makes harsh updates, as it is strongly biased towards the current mistakenly-labeled sample. Suppose at the ith step, the classifier is $\mathbf{w}_i$ and we want to make a more conservative update based on observation of $(\mathbf{x}_i, y_i)$ to a new classifier $\mathbf{w}_{i+1}$. Derive a new update method for the perceptron such that it makes the smallest difference from the previous model, that is, it minimizes $\|\mathbf{w}_{i+1} - \mathbf{w}_i\|_2$ while ensuring that $\mathbf{w}_{i+1}$ classifies the current sample correctly. You need to provide the closed form analytical equation for the update rule.

Submission

Please provide the following as part of your submission:

- Provide your answers to problems 1-4 in hardcopy. The papers need to be stapled and submitted at the beginning of class on the due date.
- You are encouraged to collaborate, but collaboration must be limited to discussion only and you need to write down / implement your solution on your own. You also need to list with whom you have discussed the HW problems.