CSCI-GA.2565-001 Machine Learning: Homework 1

Due March 5, 2019

1. **Probability**

   (a) Suppose we have samples $x_1, \ldots, x_n$ i.i.d drawn from Bernoulli(p). Find the maximum likelihood estimator of p.

   (b) Suppose we have samples $x_1, \ldots, x_n$ i.i.d drawn from uniform distribution (a, b). Find the maximum likelihood estimator of a and b.

   (c) Consider if two random variables X and Y are marginally Gaussian distributed and uncorrelated. Are X and Y independent? If so, prove it; If not, give a counterexample.

2. **Poisson Generalized Linear Models (GLM)**

   Suppose we have input-output pairs $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in N = \{0, 1, 2, 3, \ldots\}$ for $i = 1, .., n$. Here, the number of features p is greater than the number of data pairs n. Our task is to train a Poisson GLM to model the data. Assume the linear coefficients in the model is $\boldsymbol{\theta}$.

   (a) What is the log-likelihood function for Poisson GLM?

   (b) Given a test point $\mathbf{x}^*$, how do you predict with Poisson GLM?

   (c) Suppose the test point $\mathbf{x}^*$ is orthogonal to the space generated by the training data. What is the prediction ℓ_2 regularized Poisson GLM make on the test point? Prove your answer.

   (d) Use part (c) to motivate ℓ_1 regularization when the number of training points is less than the features.

3. **Very Random Forest**

   Consider building a random forest by both subsampling the data and choosing a single feature randomly. For example, consider a dataset

$\mathcal{D} = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$ where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ for $i = 1, .., n$. A tree would be constructed from data that first samples one feature index $j \leq p$ at random followed by drawing a sample of the data of size $m < n$ with replacement. From this process the subsamples are $\tilde{\mathcal{D}}_k = \{(\tilde{\mathbf{x}}_{(k_1)}, y_{(k_1)}), ..., (\tilde{\mathbf{x}}_{(k_m)}, y_{(k_m)})\}$; each $\tilde{\mathbf{x}}_{(k_l)}$ only contains the jth feature. We then build a decision tree on $\tilde{\mathcal{D}}_k$. We average many of these random trees to construct the very random forest.

In a formal way, if we have a sequence of iid random vectors $\{\theta_k\}$ controlling the mechanism of sub-sampling, i.e., θ_k generates $\tilde{\mathcal{D}}_k$ from $\mathcal{D}$, then an ensemble of trees is grown $\{h(\mathbf{x}^*; \theta_k, \mathcal{D})\}$ taking $\mathbf{x}^*$ as test input.

(a) For what true data distributions does very random forests have zero bias?

(b) On the data distributions in part (a), compare (in the sense of bias and variance) this very random forest with the traditional random forest.

4. **Alternative Losses**

Suppose we are doing binary classification on dataset

$$\{(x_1, y_1), ..., (x_n, y_n)\},$$

where $x_i \in \mathbb{R}$ and $y_i \in \{0, 1\}$ for $i = 1, .., n$. The true model for this data is linear with so that the data has $P(y_i = 1|x_i) = \frac{1}{1+e^{-\beta x_i}}$ and $P(y_i = 0|x_i) = 1 - P(y_i = 1|x_i)$, where $\beta = 3$.

Now we investigate how an extreme point would affect different estimators. Suppose a new point $(x_{n+1} = 100, y_{n+1} = 0)$ appears and we want to see how this point alters estimates by looking at derivatives of two different loss functions.

(a) Compute the derivative of the square loss for the new data point $(y_{n+1} - \sigma(\beta x_{n+1})^2$ w.r.t β at $\beta = 3$.

(b) Compute the derivative of the negative log-likelihood of logistic regression for the new data point w.r.t β at $\hat{\beta}$ where $g(\cdot)$ is the sigmoid function.

(c) Compare the results of part (a) and (b). What does it imply? Explain why. What does it tell you about how to choose losses in general?