CSE Ph.D. Qualifying Exam, Fall 2019

You should choose two areas to work on. Each area consists of four problems, and you should choose three of them. Show all your work and write in a readable way.

Data Analysis

1. **Graphical Models**

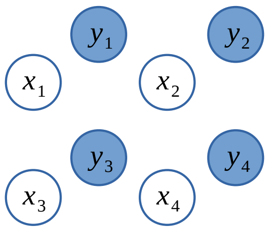
You are required to build a pairwise Markov Random Fields (MRFs) in this question. Denote the x as hidden binary variables and the y as the observed continuous variables, consider the MRFs with four hidden variables and four observed variables, we define the energy function as

$$E(x, y) = -\sum_{i,j=1}^{4} \omega_{ij} x_i x_j - \sum_{i=1}^{4} \theta_i x_i y_i.$$

Therefore, the joint distribution is

$$p(x, y) \propto \exp(-E(x, y)).$$

(a) Complete the graphical model by drawing the edges.



(b) Mark TRUE or FALSE to the following statements about conditional independence properties in the model

- $(x_1 \perp x_2 | x_3)$
- $(y_1 \perp y_2 | y_3)$
- $(y_1 \perp y_2 | x_1, x_2, x_3)$

(c) Write down the form of the normalizer, $Z(\theta, \omega)$, so that $p(x, y) = \frac{1}{Z(\theta,\omega)} \exp(-E(x, y))$ is a valid distribution.

(d) Write down the log-likelihood and its derivative w.r.t. ω_{ij}

2. **Logistic Regression**

Let us denote training set D as $\{(\mathbf{x}_i, y_i)\}_1^N$, where $y_i \in \{0, 1\}$ is the label and $\mathbf{x}_i \in \mathbf{R}^d$ is the feature vector of the i-th data point. In logistic regression we have $p(y_i = 1|\mathbf{x}_i) = \sigma(\mathbf{w}^T\mathbf{x}_i)$, where $\mathbf{w} \in \mathbf{R}^d$ is the learned coefficient vector and $\sigma(t) = \frac{1}{1+e^{-t}}$ is the sigmoid function.

1) **Batch Gradient Descent**

a) Specify the negative log-likelihood for logistic regression

b) Derive the gradient of the negative log-likelihood in terms of $\mathbf{w}$ for this setting.

2) **Stochastic Gradient Descent**

If N and d are very large, it may be prohibitively expensive to consider every patient in D before applying an update to $\mathbf{w}$. One alternative is to consider stochastic gradient descent, in which an update is applied after only considering a single data point.

a) Show the log likelihood, l, of a single data point $(\mathbf{x}_t, y_t)$.

b) Show how to update the coefficient vector $\mathbf{w}_t$ when you get a feature vector $\mathbf{x}_t$ and the label y_t at time t using $\mathbf{w}_{t-1}$ (assume learning rate η is given).

c) What is the time complexity of the update rule from **b** if $\mathbf{x}_t$ is very sparse?

d) Briefly explain the consequence of using a very large η and very small η.

e) Show how to update $\mathbf{w}_t$ with L2 regularization. That is to update $\mathbf{w}_t$ according to $l - \mu\|\mathbf{w}\|_2^2$, where μ is a constant. What's the time complexity? [5 points]

f) When you use L2 regularization, you will find each time you get a new $(\mathbf{x}_t, y_t)$ you need to update every element of vector $\mathbf{w}_t$ even if $\mathbf{x}_t$ has very few non-zero elements. Write the pseudo-code on how to update $\mathbf{w}_t$ efficiently with sparse input.

3. **Bayesian Linear Regression and Regularization**

Linear regression is a model of the form $P(y|\mathbf{x}) \sim N(\mathbf{w}^{\mathrm{T}}\mathbf{x}, \sigma^2)$ from a probabilistic point of view, where $\mathbf{w}$ is a d-dimensional vector. In ridge regression, we add an l-2 regularization term to our least squares objective function to prevent overfitting. Given data $D = \{\mathbf{x}_i, y_i\}_{i=1}^n$, our objective function for ridge regression is then:

$$J(\mathbf{w}) = \sum_{i=1}^{n}(y_i - \mathbf{w}^{\mathrm{T}}\mathbf{x}_i)^2 + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{w}.$$

We can arrive at the same objective function in a Bayesian setting, if we consider a maximum a posteriori probability (MAP) estimate and assume $\mathbf{w}$ has the prior distribution $N(0, f(\lambda, \sigma)\mathbf{I})$.

(a) Write down the posterior distribution of $\mathbf{w}$ given the data.

(b) What $f(\lambda, \sigma)$ makes this MAP estimate the same as the solution to optimizing $J(\mathbf{w})$?

4. **Random Forests**

   (a) *Random forests* is a modification over *bagging* decision trees. The random forests improves variance reduction (over bagging) by reducing correlation among trees. Briefly explain how this correlation reduction ("de-correlation") among trees is achieved when growing the trees.

   (b) Random forests are generally easy to implement and to train. It can be fit in one sequence, with cross validation performed along the way (almost identical to performing N-fold cross-validation, where N is the number of data instances), through the use of *out-of-bag* (OOB) samples. Explain why using OOB samples eliminates the need for setting aside a test set for evaluating a random forest, and how this leads to more efficient training.

   (c) List the *model hyperparameters* and *model parameters* of a random forest.

   (d) Alice and Bob are data scientists debating whether a random forest is an "interpretable" model. Alice argues that it is interpretable, while Bob argues that its interpretability is limited. Briefly discuss why they may both be correct.