SOUTH UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

Intro to Big Data Science: Assignment 3

Due Date: March 31, 2022

✏ **Exercise 1**

Log into "cookdata.cn", and enroll the course "数据科学导引". Finish the online exercise there.

✏ **Exercise 2**

We have a dataset with n records in which the i-th record has one real-valued input attribute x_i and one real-valued output response y_i. We have the following model with one unknown parameter w which we want to learn from data: $y_i \sim N(e^{wx_i}, 1)$. Note that the variance is known and equal to one.

1. Is the task of estimating w a linear regression problem or a non-linear regression problem?
2. Suppose you decide to do a maximum likelihood estimation of w. You do the math and figure out that you need w to satisfy one of the following equations. Which one?
   (A) $\sum_{i=1}^{n} x_i e^{wx_i} = \sum_{i=1}^{n} x_i y_i e^{wx_i}$
   (B) $\sum_{i=1}^{n} x_i e^{2wx_i} = \sum_{i=1}^{n} x_i y_i e^{wx_i}$
   (C) $\sum_{i=1}^{n} x_i^2 e^{wx_i} = \sum_{i=1}^{n} x_i y_i e^{wx_i}$
   (D) $\sum_{i=1}^{n} x_i^2 e^{wx_i} = \sum_{i=1}^{n} x_i y_i e^{wx_i/2}$
   (E) $\sum_{i=1}^{n} e^{wx_i} = \sum_{i=1}^{n} y_i e^{wx_i}$

✏ **Exercise 3 (Linear regression as a projection)**

Consider a multivariate liner model $\mathbf{y} = \mathbf{X}\mathbf{w} + \epsilon$ with $\mathbf{y} \in \mathbb{R}^{n\times 1}$, $\mathbf{X} \in \mathbb{R}^{n\times(d+1)}$, $\mathbf{w} \in \mathbb{R}^{(d+1)\times 1}$, and $\epsilon \in \mathbb{R}^{n\times 1}$, where $\epsilon \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, follows the normal distribution.

1. Show that the linear regression predictor is given by $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
2. Let $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, show that $\mathbf{P}$ has only 0 and 1 eigenvalues.
3. Show that $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is an unbiased estimator of $\mathbf{w}$, i.e., $\mathbf{E}(\hat{\mathbf{w}}) = \mathbf{w}$. Also show that $\mathrm{Var}(\hat{\mathbf{w}}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$. (Note that by definition, $\mathrm{Var}(\hat{\mathbf{w}}) = \mathbf{E}[(\hat{\mathbf{w}} - \mathbf{E}(\hat{\mathbf{w}}))(\hat{\mathbf{w}} - \mathbf{E}(\hat{\mathbf{w}}))^T]$).
4. Recall the definition of R^2 score: $R^2 := 1 - \frac{SS_{res}}{SS_{tot}}$, where $SS_{tot} = \sum_{i=1}^n (y_i - \bar{y})^2$, $SS_{reg} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, and $SS_{res} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$. Prove that for linear regression, $SS_{tot} = SS_{reg} + SS_{res}$. (So that R^2 score can also be defined as $R^2 = \frac{SS_{reg}}{SS_{tot}}$)

✏ **Exercise 4 (Generalized Cross-Validation)** Consider ridge regression:

$$\min_{\mathbf{w}} \left[(\mathbf{y} - \mathbf{X}\mathbf{w})^T (\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2 \right]$$

It has the solution $\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$ and prediction $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{P}\mathbf{y}$ with $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$ be the projection matrix.

1. Define the leave-one-out cross validation estimator as

$$\hat{\mathbf{w}}^{[k]} = \arg\min_{\mathbf{w}} \Big[\sum_{i=1, i\neq k}^{n} (y_i - \mathbf{x}_i^T\mathbf{w})^2 + \lambda \|\mathbf{w}\|_2^2 \Big].$$

   Show that $\hat{\mathbf{w}}^{[k]} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I} - \mathbf{x}_k\mathbf{x}_k^T)^{-1}(\mathbf{X}^T\mathbf{y} - \mathbf{x}_k y_k)$

2. Define the ordinary cross-validation (OCV) mean squared error as $V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n} (\mathbf{x}_k^T\hat{\mathbf{w}}^{[k]} - y_k)^2$. Show that $V_0(\lambda)$ can be rewritten as $V_0(\lambda) = \frac{1}{n}\sum_{k=1}^{n} \left(\frac{\hat{y}_k - y_k}{1 - p_{kk}}\right)^2$, where $\hat{y}_k = \sum_{j=1}^n p_{kj} y_j$ and p_{kj} is the (k, j)-entry of $\mathbf{P}$.

   (Hint: You may need to use the Sherman-Morrison Formula for nonsingualar matrix $\mathbf{A}$ and vectors $\mathbf{x}$ and $\mathbf{y}$ with $\mathbf{y}^T\mathbf{A}^{-1}\mathbf{x} \neq -1$: $(\mathbf{A} + \mathbf{x}\mathbf{y}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{x}\mathbf{y}^T\mathbf{A}^{-1}}{1 + \mathbf{y}^T\mathbf{A}^{-1}\mathbf{x}}$)

3. Define weights as $w_k = \left(\frac{1 - p_{kk}}{\frac{1}{n}tr(\mathbf{I} - \mathbf{P})}\right)^2$ and weighted OCV as $V(\lambda) = \frac{1}{n}\sum_{k=1}^{n} w_k(\mathbf{x}_k^T\hat{\mathbf{w}}^{[k]} - y_k)^2$. Show that $V(\lambda)$ can be written as

$$V(\lambda) = \frac{\frac{1}{n}\|(\mathbf{I} - \mathbf{A})\mathbf{y}\|^2}{\Big[1 - tr(\mathbf{P})/n\Big]^2}$$

✏ **Exercise 5 (Solving LASSO by ADMM)** The alternating direction method of multipliers (ADMM) is a very useful algorithm for solving the constrained optimization problem:

$$\min_{\theta, z} f(\boldsymbol{\theta}) + g(\mathbf{z}), \qquad \text{subject to} \quad \mathbf{A}\boldsymbol{\theta} + \mathbf{B}\mathbf{z} = \mathbf{c}.$$

The algorithm is given by using Lagrange multiplier $\mathbf{u}$ for the constraint. The detail is as follows:

1. $\boldsymbol{\theta}^{(k+1)} = \arg\min_{\boldsymbol{\theta}} L(\boldsymbol{\theta}, \mathbf{z}^{(k)}, \mathbf{u}^{(k)})$;
2. $\mathbf{z}^{(k+1)} = \arg\min_{\mathbf{z}} L(\boldsymbol{\theta}^{(k+1)}, \mathbf{z}, \mathbf{u}^{(k)})$;
3. $\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + \mathbf{A}\boldsymbol{\theta}^{(k+1)} + \mathbf{B}\mathbf{z}^{(k+1)} - \mathbf{c}$;

where L is the augmented Lagrange function defined as

$$L(\boldsymbol{\theta}, \mathbf{z}, \mathbf{u}) = f(\boldsymbol{\theta}) + g(\mathbf{z}) + \mathbf{u}^T(\mathbf{A}\boldsymbol{\theta} + \mathbf{B}\mathbf{z} - \mathbf{c}) + \frac{1}{2}\|\mathbf{A}\boldsymbol{\theta} + \mathbf{B}\mathbf{z} - \mathbf{c}\|_2^2.$$

An advantage of ADMM is that no tuning parameter such as the step size in the gradient algorithm is involved. Please write down the ADMM steps for solving LASSO problem:

$$\min_{\mathbf{w}} \left[\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda \|\mathbf{w}\|_1 \right].$$

(Hint: In order to use ADMM, you have to introduce an auxiliary variable and a suitable constraint. Please give the explicit formulae by solving "argmin" in each step of ADMM.)