CS 475 Machine Learning: Homework 2 Analytical (50 points)

Assigned: Friday, September 20, 2024

Due: Friday, October 4, 2024, 11:59 pm US/Eastern

Trevor Black (tblack20)

Instructions

We have provided this LaTeX document for turning in this homework. We give you one or more boxes to answer each question. The question to answer for each box will be noted in the title of the box. You can change the size of the box if you need more space.

Other than your name, do not type anything outside the boxes. Leave the rest of the document unchanged.

Do not change any formatting in this document, or we may be unable to grade your work. This includes, but is not limited to, the font sizes, and the spacing of text and tables. Additionally, do not add text outside of the answer boxes. Entering your answers are the only changes allowed.

We strongly recommend you review your answers in the generated PDF to ensure they appear correct. We will grade what appears in the answer boxes in the submitted PDF, NOT the original latex file.

1 Linear Regression [25 pts]

Suppose that we have a dataset of n samples $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^k$ is a feature vector, and $y_i \in \mathbb{R}$ is the corresponding target value. Each sample follows a linear model:

$$y_i = g(\mathbf{x}_i; \beta) + \epsilon_i$$

where the linear function $g(\mathbf{x}_i; \beta)$ is defined as: $g(\mathbf{x}_i; \beta) \equiv \beta_{\text{int}} + \sum_{j=1}^{k} \mathbf{x}_{ij} \cdot \beta_j$
Here:

- β_{int} is the intercept term.
- $\beta = [\beta_{\text{int}}, \beta_1, \ldots, \beta_k]$ represents the linear model parameters.
- The noise ϵ_i is independently and identically distributed as $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2\right)$, where $\mathcal{N}\left(0, \sigma^2\right)$ denotes a normal distribution with mean 0 and variance σ^2.

As we've seen in the lecture, this model can be viewed as:

$$\begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_{11} & \ldots & \mathbf{x}_{1k} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & \mathbf{x}_{n1} & \ldots & \mathbf{x}_{nk} \end{pmatrix} \times \begin{pmatrix} \beta_{\text{int}} \\ \beta_1 \\ \cdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \cdots \\ \epsilon_n \end{pmatrix}, \tag{1}$$

$$Y = X \cdot \beta + \boldsymbol{\epsilon}, \tag{2}$$

where $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2, \cdots, \epsilon_n]$ is the noise vector.

Let $\hat{\beta} = [X^\top X]^{-1} X^\top Y$ be the least squares estimator of β on the dataset $\mathcal{D}$. Note that $\hat{\beta}$ is a **random vector** due to the randomness introduced by the noise $\boldsymbol{\epsilon}$.

1. (5 pts) Show that the sum of the residuals is zero:

$$\sum_{i=1}^{n} (g(\mathbf{x}_i, \hat{\beta}) - y_i) = 0$$

$\sum_{i=1}^{n}(g(x_i, \hat{\beta}) - y_i) = \sum_{i=1}^{n}(-y_i + \beta_{int} + \sum_{j=1}^{k} x_{ij}\beta_j) = \sum_{i=1}^{n}(-y_i + \sum_{j=0}^{k} x_{ij}\beta_j)$
$= -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\sum_{j=0}^{k} x_{ij}\beta_j$
$Y = X \cdot \beta + \epsilon, Y = \{y_1, \ldots, y_n\}, X \cdot \beta = \{\sum_{j=0}^{k} x_{1j}\beta_j, \ldots, \sum_{j=0}^{k} x_{nj}\beta_j\}$
Combining these equations, we get $\{y_1, \ldots, y_n\} = \{\sum_{j=0}^{k} x_{1j}\beta_j, \ldots, \sum_{j=0}^{k} x_{nj}\beta_j\} + \epsilon$
Taking the sum of each vector, we get a new equality: $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\sum_{j=0}^{k} x_{ij}\beta_j + \epsilon$
ϵ can be set to the null vector because it is normally distributed with a mean of 0.
This gives us $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\sum_{j=0}^{k} x_{ij}\beta_j$
Plugging back into our initial problem, we get
$-\sum_{i=1}^{n} y_i + \sum_{i=1}^{n}\sum_{j=0}^{k} x_{ij}\beta_j = -\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} y_i = 0$
Therefore, $\sum_{i=1}^{n}(g(x_i, \hat{\beta}) - y_i) = 0$

2. (10 pts) Assume $\mathbb{E}[\boldsymbol{\epsilon}|X] = 0$ w.r.t the conditional distribution $P(\boldsymbol{\epsilon}|X)$, show that the least square estimator is an unbiased estimator, i.e., $\mathbb{E}_{\mathcal{D}}[\hat{\beta}] = \beta$.

$\mathbb{E}_{\mathcal{D}}[\hat{\beta}] = \mathbb{E}_{\mathcal{D}}[(X^T X)^{-1} X^T Y]$
Using the equality $Y = X\beta + \epsilon$, we can substitute Y to find
$= \mathbb{E}_{\mathcal{D}}[(X^T X)^{-1} X^T (X\beta + \epsilon)]$
By distributing $(X^T X)^{-1} X^T$ and separating the second term into its own $\mathbb{E}$, we find
$= \mathbb{E}_{\mathcal{D}}[(X^T X)^{-1} X^T X\beta] + \mathbb{E}_{\mathcal{D}}[(X^T X)^{-1} X^T \epsilon]$
$(X^T X)^{-1} X^T X = I$ (identity matrix), so
$\mathbb{E}_{\mathcal{D}}[(X^T X)^{-1} X^T X\beta] + \mathbb{E}_{\mathcal{D}}[(X^T X)^{-1} X^T \epsilon] = \mathbb{E}_{\mathcal{D}}[\beta] + \mathbb{E}_{\mathcal{D}}[(X^T X)^{-1} X^T \epsilon]$
We can also take ϵ out of the $\mathbb{E}$ and evaluate the term, as we assume $\mathbb{E}[\boldsymbol{\epsilon}|X] = 0$, giving us
$\mathbb{E}_{\mathcal{D}}[\beta] + \mathbb{E}_{\mathcal{D}}[(X^T X)^{-1} X^T \epsilon] = \mathbb{E}_{\mathcal{D}}[\beta] + 0 \cdot \mathbb{E}_{\mathcal{D}}[(X^T X)^{-1} X^T] = \mathbb{E}_{\mathcal{D}}[\beta]$
Because β is not a random variable, but a constant, we can simply extract it from the $\mathbb{E}_{\mathcal{D}}$, giving us
$\mathbb{E}_{\mathcal{D}}[\beta] = \beta$
Therefore, $\mathbb{E}_{\mathcal{D}}[\hat{\beta}] = \beta$

3. (10 pts) Assume that the covariance matrix of $\boldsymbol{\epsilon}$ is $\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right] = \sigma^2 I$.
Show that $\mathrm{Var}(\hat{\beta}) = \sigma^2 \left(X^\top X\right)^{-1}$.

As proved in the last part, $\hat{\beta} = \beta + (X^T X)^{-1} X^T \epsilon$, so
$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\beta + (X^T X)^{-1} X^T \epsilon)$
Because β is a constant, we can take it out of the Var and evaluate it to 0.
$= \mathrm{Var}((X^T X)^{-1} X^T \epsilon)$
The variance of a linear transform has the property $\mathrm{Var}(AX) = A(\mathrm{Var}(X))A^T$. Applying this where $X = \epsilon$ and $A = (X^T X)^{-1} X^T$, we get
$(X^T X)^{-1} X^T \mathrm{Var}(\epsilon) X (X^T X)^{-1} = (X^T X)^{-1} X^T \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right] X (X^T X)^{-1}$
Because $\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right] = \sigma^2 I$, we can substitute, giving us
$(X^T X)^{-1} X^T \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right] X (X^T X)^{-1} = (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1}$
And because $X^T X (X^T X)^{-1} = I$, we can find that
$= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}$
Therefore, $\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$

2 Logistic Regression [25 pts]

In binary logistic regression, we only have 2 possible classes for the outcome (i.e. $y \in \{0, 1\}$). We will generalize it to a setting where there are K classes (i.e. $y \in \{1, 2, \cdots, K\}$). Given a data sample x and a weight matrix W, we can predict the probability that x belongs to the k-th class using the softmax function with the following formula:

$$p(y = k \mid x, W) = \frac{\exp(w_k^\top x)}{\sum_{j=1}^{K} \exp(w_j^\top x)} \quad \text{for} \quad k = 1, 2, \cdots, K$$

Here:

- x is the data sample,
- W is the weight matrix, and
- $w_k^\top$ is the k-th row of W.

1. (15 pts) Since maximizing the likelihood is equivalent to minimizing the average negative log-likelihood (NLL), we want to minimize the average NLL loss to fit the model. Derive the average NLL loss for a dataset of n samples $\{(x_i, y_i)\}_{i=1}^n$, given the formula:

$$\text{NLL}(W) = -\frac{1}{n} \log \prod_{i=1}^{n} p(y_i \mid x_i, W)$$

Express the likelihood in terms of the weight parameters from the softmax function.

Average $\text{NLL}(W) = -\frac{1}{n} \log \prod_{i=1}^{n} p(y_i \mid x_i, W)$
$= -\frac{1}{n} \log \prod_{i=1}^{n} \frac{\exp(w_{y_i}^T x_i)}{\sum_{j=1}^{K} \exp(w_j^T x_i)}$
$= -\frac{1}{n} \sum_{i=1}^{n} \log \frac{\exp(w_{y_i}^T x_i)}{\sum_{j=1}^{K} \exp(w_j^T x_i)}$
$= -\frac{1}{n} \sum_{i=1}^{n} (\log \exp(w_{y_i}^T x_i) - \log \sum_{j=1}^{K} \exp(w_j^T x_i))$
$= -\frac{1}{n} \sum_{i=1}^{n} (w_{y_i}^T x_i - \log \sum_{j=1}^{K} \exp(w_j^T x_i))$

2. (10 pts) The softmax function gives us the predicted probability for each class. The true class is represented by a one-hot encoded vector — a vector of length K where all entries are 0 except for the position corresponding to the true class, which is 1.

The cross-entropy loss function is defined as:

$$\mathrm{CE}(W) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Y_{i,k}\log(\hat{Y}_{i,k})$$

where $Y_{i,k}$ is the one-hot encoded true label for sample i, and $\hat{Y}_{i,k}$ is the predicted probability for class k. Can you show that the cross-entropy loss is equivalent to the average NLL loss?

$\mathrm{CE}(W) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} Y_{i,k}\log(\hat{Y}_{i,k})$
Because the one-hot encoded $Y_{i,k}$ is 1 in the true label for sample i and is 0 everywhere else, it 'selects' the correct $k = y_i$, eliminating all other elements in the sum, giving us
$= -\frac{1}{n}\sum_{i=1}^{n}\log(\hat{Y}_{i,y_i})$
$\hat{Y}_{i,k}$ is given by the softmax function $p(y_i = k \mid x_i, W) = \frac{\exp(w_{y_i}^T x_i)}{\sum_{j=1}^{K}\exp(w_j^T x_i)}$, letting us plug in to get
$-\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp(w_{y_i}^T x_i)}{\sum_{j=1}^{K}\exp(w_j^T x_i)}$
Looking back at the last question, we can see that this is equivelant to $\mathrm{NLL}(W)$ as depicted in the third line.
Therefore, $\mathrm{CE}(W) = \mathrm{NLL}(W)$