Hong Kong University of Science and Technology
COMP 4211: Machine Learning
Fall 2021

Problem Set

Due: 12 November 2021, Friday, 11:59pm

Important Instructions:

- If an answer requires calculation or derivation, you are expected to show the steps as well in addition to the final answer.
- Your answers must be typewritten (not handwritten) properly for submission. You may use LaTeX (`http://www.latex-project.org/`), Word or other mathematical typesetting software to typeset your answers.
- Your submission must be done electronically in one PDF file via the Canvas course site.
- Late submission will be accepted but with penalty. The late penalty is deduction of one point (out of a maximum of 100 points) for every minute late after 11:59pm. Being late for a fraction of a minute is considered a full minute. For example, two points will be deducted if the submission time is 00:00:34.
- At most one NQA coupon may be used to entitle you to submit this problem set late for one day without grade penalty. You must download the file named `NQA.pdf` from Canvas and submit it by the original deadline to show that you want one coupon to be used.
- While you may discuss with your classmates on general ideas about solving the problems, your submission should be based on your own independent effort.
- In case you seek help from any person or reference source, you should state it clearly in your submission. Failure to do so is considered plagiarism which will lead to appropriate disciplinary actions (`https://acadreg.ust.hk/generalreg`).

1. **Logistic Regression** (15 points)

Recall that the output of a logistic regression model for binary classification represents the probability p that an input $\mathbf{x}$ belongs to one of the classes. Also, from statistics, the *logit* or *log-odds* of p is the logarithm of the odds ratio $p/(1-p)$, i.e., $\text{logit}(p) = \log[p/(1-p)]$.

Suppose a country has two political parties, denoted by A and B. We assume for simplicity that each voter must vote for one and only one of the two parties. The probability that a voter votes for party A or B is found to be modeled well by a logistic regression model involving a linear regression function $g(\mathbf{x})$:

$$f(\mathbf{x}) = P(\mathbf{x} \text{ votes for party A}) = \sigma(g(\mathbf{x})),$$

where $\sigma(\cdot)$ denotes the sigmoid function and

$$g(\mathbf{x}) = 0.2x_1 + 0.1x_2 + 0.3x_3 - 3.$$

By overloading the notation, we use $\mathbf{x} = (x_1, x_2, x_3)^\top$ to denote both a voter and his/her features, where x_1 is the family income (in \$10,000), x_2 is the number of years of education, and x_3 is the gender (1 for male and 0 for female).

(a) (4 points) Consider a male voter with a family income of \$50,000 and 20 years of education. According to the model, what is the probability that he will vote for party A?

(b) (4 points) Consider a female voter with a family income of \$30,000 and 12 years of education. According to the model, what is the log-odds of the probability that she will vote for party A?

(c) (4 points) Given a voter $\mathbf{x}$, show how we can use $g(\mathbf{x})$ only to predict whether the voter will vote for party A or B.

(d) (3 points) If we want to obtain a decision boundary of more general form (i.e., of more expressive power) than that in part (c), what needs to be changed in the logistic regression model?

2. **Logistic Regression** (15 points)

Let $\mathbf{x} = (x_1, x_2, \ldots, x_K)^\top$ and $\mathbf{z} = (z_1, z_2, \ldots, z_K)^\top$ be two K-dimensional vectors which are related through an activation function g_τ, i.e., $\mathbf{z} = g_\tau(\mathbf{x})$, with each dimension of $\mathbf{z}$ defined as

$$z_j = \frac{e^{x_j/\tau}}{\sum_{k=1}^{K} e^{x_k/\tau}}, \qquad 1 \le j \le K,$$

where $\tau > 0$ is called a temperature parameter.

(a) (4 points) How will $\mathbf{z}$ become as τ approaches 0?

(b) (4 points) How will $\mathbf{z}$ become as τ approaches ∞?

(c) (7 points) Derive the partial derivative of z_j with respect to x_i for all $1 \le i, j \le K$.

3. **Feedforward Neural Networks** (15 points)

Consider the 3-layer feedforward neural network shown in slide #6 of the lecture notes. Let there be $K \geq 2$ output units where each of them corresponds to one of K classes for a classification problem. The classification problem considered in the lecture notes assumes that each input $\mathbf{x}$ belongs to *one and only one* of the K classes. As such, the target output $\mathbf{y} = (y_1, \ldots, y_K)^\top$ is always represented as a one-hot vector in which only one dimension is 1 while all others are 0.

We now consider a variant of this classification problem where each input $\mathbf{x}$ belongs to *at most one* of the K classes. In other words, it is possible for $\mathbf{x}$ to belong to *none* of the K classes if $\mathbf{x}$ belongs to a class other than the K classes corresponding to the output units.

(a) (5 points) What changes, if any, should be made to the activation functions $g_j^{[1]}$, $g_k^{[2]}$, and $g_l^{[3]}$ of the original network in order to handle this variant of the classification problem? Explain your answer.

(b) (5 points) In addition, what changes, if any, should be made to the original loss function $L(\mathbf{W}; \mathcal{S})$, where $\mathcal{S} = \left\{(\mathbf{x}^{(q)}, \mathbf{y}^{(q)})\right\}_{q=1}^{N}$ is the training set?

(c) (5 points) After the network has been trained properly using $\mathcal{S}$, it can be used as a classifier to solve the classification problem described above. Given a test example $\mathbf{x}$, propose a classification rule that allows one of $K + 1$ decisions to be made about $\mathbf{x}$, where K of them correspond to the K classes and the remaining one refers to 'none of the above' or a 'reject' option indicating that $\mathbf{x}$ does not belong to any of the K classes.

4. **Convolutional Neural Networks** (20 points)

For each convolutional layer, let (n_i, n_o, h, w, s) be a concise way to represent its configuration which corresponds to n_i input channels, n_o output channels (a.k.a. feature maps), one convolution kernel of height h and width w for each feature map, and a stride of s.

Let us assume that no zero padding is applied.

(a) Suppose we have a convolutional layer with configuration $(128, 32, 7, 7, 2)$ and the image in each input channel is of size 261×261.

i. (3 points) What is the size of each feature map in the convolutional layer?

ii. (3 points) How many parameters are there in the convolutional layer?

(b) To reduce the number of parameters, we now add a new convolutional layer before the original one. The new convolutional layer has configuration $(128, 16, 1, 1, 1)$ and the configuration of the original convolutional layer is modified to $(16, 32, 7, 7, 2)$.

i. (3 points) What is the size of each feature map in the new convolutional layer?

ii. (3 points) How many parameters are there in the new convolutional layer?

iii. (3 points) What is the size of each feature map in the original convolutional layer after its configuration is modified?

iv. (3 points) How many parameters are there in the original convolutional layer after modification?

v. (2 points) By introducing an additional convolutional layer with 1×1 kernels, the number of parameters is reduced. What is the saving in percentage?

5. **Principal Component Analysis** (10 points)

Let $\mathcal{S} = \{\mathbf{x}^{(\ell)} \in \mathbb{R}^2\}_{\ell=1}^{N}$ be a set of N data points in a two-dimensional input space. The sample covariance matrix is:

$$\mathbf{C} = \begin{bmatrix} 7 & \sqrt{5} \\ \sqrt{5} & 3 \end{bmatrix}.$$

We perform principal component analysis (PCA) on $\mathcal{S}$.

(a) (3 points) If we use two principal components obtained by PCA to project $\mathcal{S}$ linearly onto another space spanned by the principal components, what is the maximum percentage of total variance that can be explained by the two principal components together? Explain your answer.

(b) (7 points) If now we use only one principal component obtained by PCA to project $\mathcal{S}$ linearly onto another space spanned by the principal component, what is the maximum percentage of total variance that can be explained by the principal component? Explain your answer.

6. **Clustering – Partitional Clustering** (10 points)

(a) (6 points) Given a set of data points $\mathcal{S} = \left\{\mathbf{x}^{(\ell)}\right\}_{\ell=1}^{N}$, the k-means clustering algorithm seeks to find k reference vectors $\{\mathbf{m}_i\}_{i=1}^{k}$ that minimize the sum of squared errors defined as follows:

$$E(\{\mathbf{m}_i\}_{i=1}^{k}; \mathcal{S}) = \sum_{\ell=1}^{N} \sum_{i=1}^{k} b_i^{(\ell)} \left\|\mathbf{x}^{(\ell)} - \mathbf{m}_i\right\|^2,$$

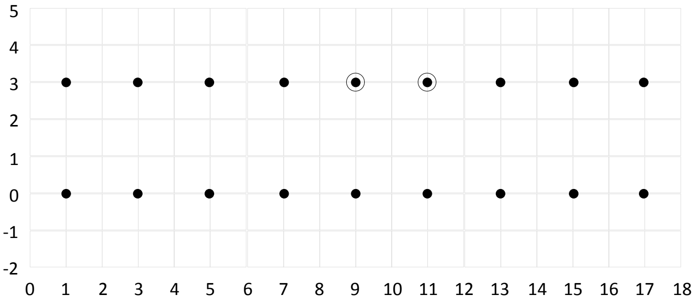
where

$$b_i^{(\ell)} = \begin{cases} 1 & \text{if } i = \arg\min_j \left\|\mathbf{x}^{(\ell)} - \mathbf{m}_j\right\| \\ 0 & \text{otherwise.} \end{cases}$$

However, there is no guarantee that k-means finds the global optimum because it is only a local optimization algorithm which alternates between optimizing $\left\{b_i^{(\ell)}\right\}$ and $\{\mathbf{m}_i\}$ by fixing one set of variables at a time.

When the set $\left\{b_i^{(\ell)}\right\}$ is held fixed, show that the update equation for each $\mathbf{m}_i$ minimizes the objective function $E(\{\mathbf{m}_i\}_{i=1}^{k}; \mathcal{S})$.

(b) (4 points) Shown below are 18 data points represented as black dots lying on the 2D plane. We apply the k-means clustering algorithm to these data points with two reference vectors, shown as circles, initialized at $\mathbf{m}_1 = (9, 3)$ and $\mathbf{m}_2 = (11, 3)$. What will be the locations of $\mathbf{m}_1$ and $\mathbf{m}_2$ after k-means converges? How many data points will be assigned to each of the two clusters formed?



7. **Clustering – Hierarchical Clustering** (15 points)

We apply agglomerative clustering to a set of one-dimensional data points

$$\mathcal{S} = \{2, 4, 5, 20, 25, 39, 43, 44\}.$$

(a) (6 points) Using single-link proximity with the Euclidean distance measure, draw the resulting dendrogram (i.e., hierarchical tree diagram) for the data set. The vertical axis should indicate the distance between clusters. From the dendrogram, identify the three topmost-level clusters by listing the three corresponding subsets of data points.

(b) (6 points) Repeat part (a) by using complete-link proximity instead.

(c) (3 points) Do the two dendrograms have the same hierarchical structure? Are the two partitions with three subsets each the same?