CAP6610 Machine Learning, Spring 2022

Midterm 1 Solution

1. (15 points) *Fitting a piecewise linear function to data.* We are given some samples $(x_i, y_i), i = 1, \ldots, n$, with x_i, y_i scalars, with a function of the form

$$\hat{y} = f(x) = \theta_1 + \theta_2 \max(0, x) + \theta_3 \min(0, x),$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ contains the parameters. The function is affine for $x < 0$ and also for $x > 0$, and it is continuous at $x = 0$ with value θ_1. Such a function is called *piecewise affine*, or more commonly *piecewise linear*, with a single knot or kink point at 0. We choose the parameters using the least squares regressionn on the given data points.

Consider the specific data set of (x, y) pairs

$$(-2, 3), (-1, 1), (0, 1), (1, 3), (2, 2)$$

Give the matrix $\boldsymbol{\Phi}$ and vector $\boldsymbol{\psi}$ for which the sum of the squares of the fitting errors is equal to $\|\boldsymbol{\Phi\theta} - \boldsymbol{\psi}\|^2$. We want the explicit numerical values of $\boldsymbol{\Phi}$ and $\boldsymbol{\psi}$. Then use any programming language to solve this least squares problem and give the numerical values of the optimal $\boldsymbol{\theta}$.

Solution.

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & 0 & -2 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 0 \end{bmatrix}, \qquad \boldsymbol{\psi} = \begin{bmatrix} 3 \\ 1 \\ 1 \\ 3 \\ 2 \end{bmatrix}.$$

Solving the least squares gives

$$\boldsymbol{\theta} = \begin{bmatrix} 1.1429 \\ 0.7143 \\ -0.7143 \end{bmatrix}.$$

In the initial version the model is $\hat{y} = f(x) = \theta_1 + \theta_2 \max(0, x) + \theta_3 \min(0, x)$. The corresponding $\boldsymbol{\Phi}$ matrix is

$$\begin{bmatrix} -2 & 0 & -2 \\ -1 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 0 \end{bmatrix}.$$

This matrix is rank-deficient since the sum of the last two columns equals to the first one. Nevertheless, if you solve it using the backslash command `Phi\psi`, the solution is $(0, 1.4, -1.4)$. You will not lose points for solving it using this setup.

2. (15 points) Consider the document classification problem. The training data is a set of documents with their correct labels. Suppose the docs are categorizated into k classes, and y_i denotes the class that doc i belongs to. Each doc is transformed into the "bag-of-words" representation, say $\boldsymbol{x}_i$ for the ith doc, meaning the jth entry of $\boldsymbol{x}_i$ represents the number of times term j appears in doc i. We want to estimate the probability $p(y|\boldsymbol{x})$ in order to design a document classifier. We use a generative approach: according to Bayes' rule, $p(y|\boldsymbol{x})$ is proportional to $p(y)p(\boldsymbol{x}|y)$:

   (a) Describe how to estimate $p(y)$

   (b) Assume $p(\boldsymbol{x}|y)$ follows a multinomial distribution; for each class c, the parameter for the multinomial distribution is a nonnegative vector $\boldsymbol{p}_c$ that sums to one. Describe how to estimate $\boldsymbol{p}_c$.

   (c) Write the resulting classifier in the form $\widehat{y} = \max_c(\boldsymbol{w}_c^\top \boldsymbol{x} + \beta_c)$. Explain how to obtain $\boldsymbol{w}_c$ and β_c from $p(y)$ and $p(\boldsymbol{x}|y)$.

   Solution.

   (a) $p(y)$ is simply obtained from counting the number of documents of a specific class divided by the total number of docs, i.e.,
   $$p(c) = \pi_c = \frac{n_c}{\sum_{j=1}^k n_j}.$$

   (b) The vector $\boldsymbol{p}_c$ represents the probability of each term appearing in a doc from class c. A "bag-of-words" representation of a document is a histogram of terms, and we have multiple of them from the same class. The maximum likelihood estimate amounts to combining all of them into a single histogram and normalize:
   $$\boldsymbol{p}_c = \frac{1}{\sum_{i \in \text{class } c} \boldsymbol{1}^\top \boldsymbol{x}_i} \sum_{i \in \text{class } c} \boldsymbol{x}_i.$$

   (c) The probability that a new document $\boldsymbol{x}$ belongs to class c is proportional to
   $$\pi_c \prod_{j=1}^m p_{cj}^{x_j},$$
   where the term that involve the factorials are the same for all classes, therefore inconsequential in making the predictions. Taking the log of it resulting in the following classifier:
   $$\begin{aligned} \widehat{y} &= \arg\max_c \left(\log \pi_c + \sum_{j=1}^m x_j \log p_{cj} \right) \\ &= \arg\max_c (\boldsymbol{w}_c^\top \boldsymbol{x} + \beta_c), \end{aligned}$$
   where
   $$\boldsymbol{w}_c(j) = \log p_{cj}, \qquad \beta_c = \log \pi_c.$$

3. (20 points) *Some sets of probability distributions.* Let x be a real-valued random variable with sample space $\{a_1, \ldots, a_k\}$ where $a_1 \le a_2 \le \cdots \le a_k$. This can be view as a categorical random variable with each catogory assigned a real value. Let $\Pr[x = a_i] = p_i$, then the vector $\boldsymbol{p}$ satisfies $\boldsymbol{p} \ge 0$ and $\boldsymbol{1}^\top \boldsymbol{p} = 1$, i.e., it lies in the probability simplex Δ. Which of the following conditions are convex in $\boldsymbol{p}$? (That is, defines a set of $\boldsymbol{p}$ that is a convex set.) Circle the correct answers. You do not need to justify your responses.

   (a) $\alpha \le \mathrm{E}[f(x)] \le \beta$, where $\mathrm{E}[f(x)]$ is the expected value of $f(x)$, i.e., $\mathrm{E}[f(x)] = \sum_{i=1}^k p_i f(a_i)$. (The function $f : \mathbb{R} \to \mathbb{R}$ is given.)

   (b) $\Pr[x > \alpha] \le \beta$.

(c) $\mathrm{E}[x^3] \leq \alpha\,\mathrm{E}[x]$.

(d) $\mathrm{var}(x) \leq \alpha$.

(e) $\mathrm{var}(x) \geq \alpha$.

Solution.

(a) $\mathrm{E}[f(x)] = \sum_{i=1}^k p_i f(a_i)$ is a linear function of $\boldsymbol{p}$ if $\boldsymbol{a}$ and $f(\cdot)$ are given. Therefore $\alpha \leq \mathrm{E}[f(x)] \leq \beta$ is the intersection of two half-spaces, which is a convex set.

(b) Define vector $\boldsymbol{b} \in \mathbb{R}^k$ as

$$b_i = \begin{cases} 1 & a_i > \alpha \\ 0 & \text{otherwise.} \end{cases}$$

Then the inequality is equivalent to $\boldsymbol{b}^\top \boldsymbol{p} \leq \beta$, which defines a half-space and thus convex.

(c) Define vector $\boldsymbol{c} \in \mathbb{R}^k$ with $c_i = a_i^3 - \alpha a_i$, then the inequality is equivalent to $\boldsymbol{c}^\top \boldsymbol{p} \leq 0$. This is another half-space and thus convex.

(d) The inequality translates to $\mathrm{var}(x) = \sum_{i=1}^k p_i a_i^2 - (\sum_{i=1}^k p_i a_i)^2 \leq \alpha$, which is a concave function less than constant. This is in general not convex. As an example, we can take $k = 2$, $a_1 = 0, a_2 = 1$, and $\alpha = 1/5$. $p = (1, 0)$ and $p = (0, 1)$ are two feasible points, but the convex combination $p = (1/2, 1/2)$ is not.

(e) This is a concave function *bigger than or equal to* constant, which defines a convex set.

4. (17 points) *Maximum likelihood estimation for exponential family.* A probability distribution or density on a set $\mathcal{X}$, parameterized by $\boldsymbol{\theta} \in \mathbb{R}^m$, is called an *exponential family* if it has the form

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = a(\boldsymbol{\theta}) \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})),$$

for $\boldsymbol{x} \in \mathcal{X}$, where $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^m$, and $a(\boldsymbol{\theta})$ is a normalization function. Here we interpret as a density function when $\mathcal{X}$ is a continuous set, and a probability distribution if $\mathcal{X}$ is discrete. Thus we have

$$a(\boldsymbol{\theta}) = \left(\int_{\mathcal{X}} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})) d\boldsymbol{x} \right)^{-1}$$

when $p(\boldsymbol{x}; \boldsymbol{\theta})$ is a density, and

$$a(\boldsymbol{\theta}) = \left(\sum_{\boldsymbol{x} \in \mathcal{X}} \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x})) \right)^{-1}$$

when $p(\boldsymbol{x}; \boldsymbol{\theta})$ represents a distribution. We consider only values of $\boldsymbol{\theta}$ for which the integral or sum above is finite. Many families of distributions have this form, for appropriate choice of the parameter $\boldsymbol{\theta}$ and function $\boldsymbol{\phi}$.

Show that for any $\boldsymbol{x} \in \mathcal{X}$, the log-likelihood function $\log p(\boldsymbol{x}; \boldsymbol{\theta})$ is concave in $\boldsymbol{\theta}$. This means that maximum-likelihood estimation for an exponential family leads to a convex optimization problem. You don't have to give a formal proof of concavity of $\log p(\boldsymbol{x}; \boldsymbol{\theta})$ in the general case: You can just consider the case when $\mathcal{X}$ is finite, and state that the other cases (discrete but infinite $\mathcal{X}$, continuous $\mathcal{X}$) can be handled by taking limits of finite sums.

Solution. Suppose $\mathcal{X}$ is a finite set $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$, then the maximum likelihood estimation of $\boldsymbol{\theta}$ leads to the formulation

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \log \left(\sum_{c=1}^k \exp(\boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_c)) \right) - \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}).$$

This is a convex optimization problem because the first term is the log-sum-exp function $\log(\sum_{c=1}^k \exp(z_c))$ composing with an affine function $z_c = \boldsymbol{\theta}^\top \boldsymbol{\phi}(\boldsymbol{x}_c), c = 1, \ldots, k$, which is convex, and the second term is

linear. Notice the similarity between this formulation and the multi-class logistic classification (cross-entropy) formulation.

The other cases (discrete but infinite $\mathcal{X}$, continuous $\mathcal{X}$) can be handled by taking limits of finite sums.

5. (17 points) *Fitting with censored data.* In some experiments there are two kinds of measurements or data available: The usual ones, in which you get a number (say), and censored data, in which you don't get the specific number, but are told something about it, such as a lower bound. A classic example is a study of lifetimes of a set of subjects (say, laboratory mice). For those who have died by the end of data collection, we get the lifetime. For those who have not died by the end of data collection, we do not have the lifetime, but we do have a lower bound, i.e., the length of the study. These are the censored data values.

   We wish to fit a set of data points $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, with $\boldsymbol{x}_i \in \mathbb{R}^m$ and $y \in \mathbb{R}$, using a linear model of the form $y \approx \boldsymbol{x}^\top \boldsymbol{w}$. The vector $\boldsymbol{w} \in \mathbb{R}^m$ is the model parameter, which we wish to choose. We will use a least-squares criterion, i.e., choose $\boldsymbol{w}$ to minimize

$$\sum_{i=1}^{n} \; (y_i - \boldsymbol{x}_i^\top \boldsymbol{w})^2.$$

   Here is the tricky part: some of the values of y_i are censored; for these entries, we have only a (given) lower bound. We will re-order the data so that $y_1, \ldots, y_p$ are given (i.e., uncensored), while $y_{p+1}, \ldots, y_n$ are all censored, i.e., unknown, but larger than D, a given number. All the values of $\boldsymbol{x}_i$ are known.

   Formulate an optimization problem to find $\boldsymbol{w}$ (the model parameter) and $y_{p+1}, \ldots, y_n$ (the censored data values). Be sure to clarify the three essential components of an optimization problem: the loss function to be minimized, the optimization variables, and constraints (if any). Is it a convex problem?

   Solution.

$$\begin{array}{ll} \underset{\boldsymbol{w}, y_{p+1}, \ldots, y_n}{\text{minimize}} & \displaystyle\sum_{i=1}^{n} \; (y_i - \boldsymbol{x}_i^\top \boldsymbol{w})^2 \\ \text{subject to} & y_i \geq D, i = p+1, \ldots, n. \end{array}$$

   It is a convex optimization problem. The loss function is convex w.r.t. $y_{p+1}, \ldots, y_n$ and $\boldsymbol{w}$. The constraints are linear.

6. (16 points) *Poisson regression.* Consider a regression problem with training data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$. Again our predictor takes the form $\widehat{y} = \boldsymbol{\phi}^\top \boldsymbol{\theta}$, where $\boldsymbol{\phi}$ is a feature vector obtained from $\boldsymbol{x}$, and the coefficients $\boldsymbol{\theta}$ is learned from a supervised learning model.

   In many cases the target output y is always a positive integer, so it makes sense to assume that it follows a Poisson distribution. Specifically, we assume that the conditional probability $p(y|\boldsymbol{x})$ follows a Poisson distribution with parameter $\lambda = e^{\boldsymbol{\phi}^\top \boldsymbol{\theta}}$. Recall that the Poisson pmf with parameter λ is

$$p(y) = \frac{\lambda^y e^{-\lambda}}{y!}.$$

   Formulate an optimization problem to learn $\boldsymbol{\theta}$. Is it a convex optimization problem?
   Hint. We will maximize the log-likelihood of the training data, assuming the samples are independent.

   Solutoin. According to this model,

$$p(y|\boldsymbol{x}) = \frac{e^{y\boldsymbol{\phi}^\top \boldsymbol{\theta}} e^{-e^{\boldsymbol{\phi}^\top \boldsymbol{\theta}}}}{y!}.$$

Its log-likelihood of the entire training data is

$$\sum_{i=1}^{n} \left(y_i \boldsymbol{\phi}_i^\top \boldsymbol{\theta} - e^{\boldsymbol{\phi}_i^\top \boldsymbol{\theta}} - \log(y_i!) \right).$$

This leads to the following optimization problem by maximizing the log-likelihood (notice that the last term does not depend on $\boldsymbol{\theta}$ and thus irrelevant)

$$\underset{\boldsymbol{\theta}}{\text{minimize}} \quad \sum_{i=1}^{n} \left(e^{\boldsymbol{\phi}_i^\top \boldsymbol{\theta}} - y_i \boldsymbol{\phi}_i^\top \boldsymbol{\theta} \right).$$

This is a convex optimization problem because e^x is convex in x and the objective function is a summation over affine compositions of this function plus a linear term.