

Problem Set 6, Oct 27, 2022 (Theory Questions)

1 Convexity

Recall that we say that a function f is *convex* if the domain of f is a convex set and

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y), \text{ for all } x, y \text{ in the domain of } f, 0 \leq \theta \leq 1.$$

And *strictly convex* if

$$f(\theta x + (1 - \theta)y) < \theta f(x) + (1 - \theta)f(y), \text{ for all } x \neq y \text{ in the domain of } f, 0 < \theta < 1.$$

Prove the following assertions.

1. The affine function $f(x) = ax + b$ is convex, where a, b and x are scalars.
2. If multiple functions $f_n(x)$ are convex over a fixed domain, then their sum $g(x) = \sum_n f_n(x)$ is convex over the same domain.
3. Take $f, g : \mathbb{R} \rightarrow \mathbb{R}$ to be convex functions and g to be increasing. Then the function $g \circ f$ defined as $(g \circ f)(x) = g(f(x))$ is also convex.
Note: A function g is increasing if $a \geq b \Leftrightarrow g(a) \geq g(b)$. An example of a convex and increasing function is $\exp(x), x \in \mathbb{R}$.
4. If $f : \mathbb{R} \rightarrow \mathbb{R}$ is convex, then $g : \mathbb{R}^D \rightarrow \mathbb{R}$, where $g(x) := f(w^\top x + b)$, is also convex. Here, w is a constant vector in \mathbb{R}^D , b is a constant in \mathbb{R} and $x \in \mathbb{R}^D$.
5. Let $f : \mathbb{R}^D \rightarrow \mathbb{R}$ be strictly convex. Let $x^* \in \mathbb{R}^D$ be a global minimizer of f . Show that this global minimizer is unique. Hint: Do a proof by contradiction.

2 Extension of Logistic Regression to Multi-Class Classification

Suppose we have a classification dataset with N data example pairs $\{x_n, y_n\}$, $n \in [1, N]$, and y_n is a categorical variable over K categories, $y_n \in \{1, 2, \dots, K\}$. We wish to fit a linear model in a similar spirit to logistic regression, and we will use the softmax function to link the linear inputs to the categorical output, instead of the logistic function.

We will have K sets of parameters w_k , and define $\eta_{nk} = w_k^\top x_n$ and compute the probability distribution of the output as follows,

$$\mathbb{P}[y_n = k \mid x_n, w_1, \dots, w_K] = \frac{\exp(\eta_{nk})}{\sum_{j=1}^K \exp(\eta_{nj})}.$$

Note that we indeed have a probability distribution, as $\sum_{k=1}^K \mathbb{P}[y_n = k \mid x_n, w_1, \dots, w_K] = 1$. To make the model *identifiable*, we will fix w_K to $\mathbf{0}$, which means we have $K-1$ sets of parameters to learn. As in logistic regression, we will assume that each y_n is i.i.d., i.e.,

$$\mathbb{P}[y \mid \mathbf{X}, w_1, \dots, w_K] = \prod_{n=1}^N \mathbb{P}[y_n \mid x_n, w_1, \dots, w_K].$$

1. Derive the log-likelihood for this model.

Hint: It might be helpful to use the indicator function $1_{y_n=k}$, that is equal to one if $y_n = k$ and 0 otherwise.

2. Derive the gradient with respect to each \mathbf{w}_k .
3. Show that the negative of the log-likelihood is jointly convex in $\mathbf{w}_1, \dots, \mathbf{w}_K$.
Hint: you can use Hölder's inequality:

$$\sum_k |x_k y_k| \leq \left(\sum_k |x_k|^p \right)^{\frac{1}{p}} \left(\sum_k |y_k|^q \right)^{\frac{1}{q}},$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

3 Mixture of Linear Regression

If you have a regression dataset with two or more distinct clusters, a mixture of linear regression models is preferred over just one linear regression model.

Consider a regression dataset with N pairs $\{y_n, \mathbf{x}_n\}$. Similar to Gaussian mixture model (GMM), let $r_n \in \{1, 2, \dots, K\}$ index the mixture component. The distribution of the output y_n under the k^{th} linear model is defined as follows:

$$p(y_n | \mathbf{x}_n, r_n = k, \mathbf{w}) := \mathcal{N}(y_n | \mathbf{w}_k^\top \tilde{\mathbf{x}}_n, \sigma^2)$$

Here, \mathbf{w}_k is the regression parameter vector for the k^{th} model with \mathbf{w} being a vector containing all \mathbf{w}_k . Also, denote $\tilde{\mathbf{x}}_n := [1, \mathbf{x}_n^\top]^\top$.

1. Define \mathbf{r}_n to be a binary vector of length K such that all the entries are 0 except the k^{th} entry, i.e., $r_{nk} = 1$, implying that \mathbf{x}_n is assigned to the k^{th} mixture. Rewrite the likelihood $p(y_n | \mathbf{x}_n, \mathbf{w}, \mathbf{r}_n)$ in terms of r_{nk} .
2. Write the expression for the joint distribution $p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \mathbf{r})$ where \mathbf{r} is the set of all $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$.
3. Assume that r_n follows a multinomial distribution $p(r_n = k | \boldsymbol{\pi}) = \pi_k$, with $\boldsymbol{\pi} = [\pi_1, \pi_2, \dots, \pi_K]$. Derive the marginal distribution $p(y_n | \mathbf{x}_n, \mathbf{w}, \boldsymbol{\pi})$ obtained after marginalizing r_n out.
4. Write the expression for the maximum likelihood estimator $\mathcal{L}(\mathbf{w}, \boldsymbol{\pi}) := -\log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \boldsymbol{\pi})$ in terms of data \mathbf{y} and \mathbf{X} , and parameters \mathbf{w} and $\boldsymbol{\pi}$.
5. (a) Is \mathcal{L} jointly convex with respect to \mathbf{w} and $\boldsymbol{\pi}$?
(b) Is the model identifiable? That is, can the MLE estimator return the true parameters \mathbf{w} and $\boldsymbol{\pi}$, assuming we have infinitely many samples.
Prove your answers.