Labs
Machine Learning Course
Fall 2020

EPFL
School of Computer and Communication Sciences
Nicolas Flammarion & Martin Jaggi
www.epfl.ch/labs/mlo/machine-learning-cs-433

Problem Set 6, Oct 20, 2020 (Theory Questions)

1 Convexity

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

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

And *strictly convex* if

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

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 $f_n(\boldsymbol{x})$ are convex over a fixed domain, then their sum $g(\boldsymbol{x}) = \sum_n f_n(\boldsymbol{x})$ is convex over the same domain.
3. Take $f, g : \mathbb{R} \to \mathbb{R}$ 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 $a \ge b \Leftrightarrow g(a) \ge g(b)$. An example of a convex and increasing function is $\exp(x), x \in \mathbb{R}$.
4. If $f : \mathbb{R} \to \mathbb{R}$ is convex, then $g : \mathbb{R}^D \to \mathbb{R}$, where $g(\boldsymbol{x}) := f(\boldsymbol{w}^\top \boldsymbol{x} + b)$, is also convex. Here, $\boldsymbol{w}$ is a constant vector in $\mathbb{R}^D$, b is a constant in $\mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^D$.
5. Let $f : \mathbb{R}^D \to \mathbb{R}$ be strictly convex. Let $\boldsymbol{x}^\star \in \mathbb{R}^D$ 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 $\{\boldsymbol{x}_n, y_n\}$, $n \in [1, N]$, and y_n is a categorical variable over K categories, $y_n \in \{1, 2, ..., K\}$. 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 $\boldsymbol{w}_k$, and define $\eta_{nk} = \boldsymbol{w}_k^\top \boldsymbol{x}_n$ and compute the probability distribution of the output as follows,

$$\mathbb{P}[y_n = k \,|\, \boldsymbol{x}_n, \boldsymbol{w}_1, ..., \boldsymbol{w}_K] = \frac{\exp(\eta_{nk})}{\sum_{j=1}^K \exp(\eta_{nj})}.$$

Note that we indeed have a probability distribution, as $\sum_{k=1}^K \mathbb{P}[y_n = k \,|\, \boldsymbol{x}_n, \boldsymbol{w}_1, ..., \boldsymbol{w}_K] = 1$. To make the model *identifiable*, we will fix $\boldsymbol{w}_K$ 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.,

$$\mathbb{P}[\boldsymbol{y} \,|\, \mathbf{X}, \boldsymbol{w}_1, ..., \boldsymbol{w}_K] = \prod_{n=1}^N \mathbb{P}[y_n \,|\, \boldsymbol{x}_n, \boldsymbol{w}_1, ..., \boldsymbol{w}_K].$$

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 $\boldsymbol{w}_k$.

3. Show that the negative of the log-likelihood is convex with respect to $\boldsymbol{w}_k$.
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 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 $\{y_n, \boldsymbol{x}_n\}$. Similar to Gaussian mixture model (GMM), let $r_n \in \{1, 2, \ldots, K\}$ index the mixture component. Distribution of the output y_n under the k^{th} linear model is defined as follows:

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

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

1. Define $\boldsymbol{r}_n$ to be a binary vector of length K such that all the entries are 0 except a k^{th} entry, i.e., $r_{nk} = 1$, implying that $\boldsymbol{x}_n$ is assigned to the k^{th} mixture. Rewrite the likelihood $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{r}_n)$ in terms of r_{nk}.

2. Write the expression for the joint distribution $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{r})$ where $\boldsymbol{r}$ is the set of all $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N$.

3. Assume that r_n follows a multinomial distribution $p(r_n = k|\boldsymbol{\pi}) = \pi_k$, with $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_K]$. Derive the marginal distribution $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{\pi})$ obtained after marginalizing r_n out.

4. Write the expression for the maximum likelihood estimator $\mathcal{L}(\boldsymbol{w}, \boldsymbol{\pi}) := -\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{\pi})$ in terms of data $\boldsymbol{y}$ and $\boldsymbol{X}$, and parameters $\boldsymbol{w}$ and $\boldsymbol{\pi}$.

5. (a) Is $\mathcal{L}$ jointly-convex with respect to $\boldsymbol{w}$ and $\boldsymbol{\pi}$?
(b) Is the model identifiable? I.e. does the MLE estimator always gives the true parameters $\boldsymbol{w}$ and $\boldsymbol{\pi}$ asymptotically if the number of samples N is going to infinity.
Prove your answers.