Theory problems
Machine Learning Course
Fall 2016

EPFL
School of Computer and Communication Sciences
Martin Jaggi & Rüdiger Urbanke
mlo.epfl.ch/page-136795.html
epfmlcourse@gmail.com

Problem Set 7, Nov 3, 2016 (Solutions to Theory Questions)

1 Convexity

1. We need to check that

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$$

for all $x, y \in \mathbb{R}$ and $\theta \in [0,1]$. Since the function is linear, we get an equality and the expression is equal to

$$a(\theta x + (1-\theta)y) = b.$$

2. For any elements x, y in the common fixed domain we have that

$$\begin{aligned} g(\theta x + (1-\theta)y)) &= \sum_i f_i(\theta x + (1-\theta)y) \\ &\leq \sum_i [\theta f_i(x) + (1-\theta)f_i(y)] \\ &= \theta \sum_i f_i(x) + (1-\theta)\sum_i f_i(y) \\ &= \theta g(x) + (1-\theta)g(y). \end{aligned}$$

3. Recall: In one dimension, a function is convex if and only if its second derivative is non-negative.
Let $h(x) = g(f(x))$. We have that

$$\begin{aligned} h'(x) &= g'(f(x))f'(x), \\ h''(x) &= g''(f(x))(f'(x))^2 + g'(f(x))f''(x). \end{aligned}$$

- Since g is convex, $g'' \geq 0$.
- Since g is increasing, $g' \geq 0$.
- Since f is convex, $f'' \geq 0$.

Combining these three observations, we see that $h'' \geq 0$, i.e., h is convex.

4. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two elements in the domain. Let $x = \boldsymbol{w}^\top \boldsymbol{x} + b$ and $y = \boldsymbol{w}^\top \boldsymbol{y} + b$. Let $\theta \in [0,1]$. We need to show that

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y),$$

which follows since by assumption f was convex.

5. Assume that it has two global minima at $x^\star$ and $y^\star$. Let $z^\star = (x^\star + y^\star)/2$. Then, since f is strictly convex, we have $f(z^\star) < \frac{1}{2}(f(x^\star) + f(y^\star)) = f(x^\star) = f(y^\star)$, which contradicts the global minimality of the two points $x^\star$ and $y^\star$.

2 Extension of Logistic Regression to Multi-Class Classification

1. We will use $\mathbf{W} = \mathbf{w}_1, ..., \mathbf{w}_K$ to avoid heavy notation. We have that

$$\log \mathbb{P}[\hat{\mathbf{y}} = \mathbf{y}|\mathbf{X}, \mathbf{W}] = \log \prod_{n=1}^{N} \mathbb{P}[\hat{y}_n = y_n|\mathbf{x}_n, \mathbf{W}]$$

Where $\hat{\mathbf{y}}$ are our predictions and $\mathbf{y}$ represent the ground truth for our samples. We can rewrite the equation as follow, dividing the samples in groups based on their class.

$$\log \mathbb{P}[\hat{\mathbf{y}} = \mathbf{y}|\mathbf{X}, \mathbf{W}] = \log \prod_{n:y_n=1} \mathbb{P}[\hat{y}_n = 1|\mathbf{x}_n, \mathbf{W}] ... \prod_{n:y_n=K} \mathbb{P}[\hat{y}_n = K|\mathbf{x}_n, \mathbf{W}]$$

We introduce the following notation to simplify the expression. Let $1_{y_n=k}$ be the indicator function for $y_n = k$, i.e., it is equal to one if $y_n = k$ and 0 otherwise. Notice that we can write that

$$\mathbb{P}[\hat{y}_n = k|\mathbf{x}_n, \mathbf{W}] = \prod_{j=1}^{K} \mathbb{P}[\hat{y}_n = j|\mathbf{x}_n, \mathbf{W}]^{1_{y_n=j}},$$

as $\mathbb{P}[\hat{y}_n = j|\mathbf{x}_n, \mathbf{W}]^{1_{y_n=j}}$ is 1 when $j \neq k$ (elevating to 0), whereas $\mathbb{P}[\hat{y}_n = k|\mathbf{x}_n, \mathbf{W}]$ is left unchanged.

$$\begin{aligned}
\log \mathbb{P}[\hat{\mathbf{y}} = \mathbf{y}|\mathbf{X}, \mathbf{W}] &= \log \prod_{k=1}^{K} \prod_{n=1}^{N} \mathbb{P}[y_n = k|\mathbf{x}_n, \mathbf{W}]^{1_{y_n=k}} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \log \mathbb{P}[y_n = k|\mathbf{x}_n, \mathbf{W}] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \left[\mathbf{w}_k^\top \mathbf{x}_n - \log \sum_{j=1}^{K} \exp(\mathbf{w}_j^\top \mathbf{x}_n) \right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \mathbf{w}_k^\top \mathbf{x}_n - \sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \log \sum_{j=1}^{K} \exp(\mathbf{w}_j^\top \mathbf{x}_n) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \mathbf{w}_k^\top \mathbf{x}_n - \sum_{n=1}^{N} \log \sum_{k=1}^{k} \exp(\mathbf{w}_k^\top \mathbf{x}_n).
\end{aligned}$$

The last step is obtained by $\sum_{k=1}^{K} 1_{y_n=k} = 1$.

2. We get

$$\frac{\partial \log \mathbb{P}[\mathbf{y}|\mathbf{X}, \mathbf{W}]}{\partial \mathbf{w}_k} = \sum_{n=1}^{N} 1_{y_n=k} \mathbf{x}_n - \sum_{n=1}^{N} \text{softmax}(\eta, k) \mathbf{x}_n.$$

Where $\text{softmax}(\eta, k) = \frac{\exp(\eta_k)}{\sum_{i=1}^{K} \exp(\eta_i)}$.

3. The negative of the log-likelihood is

$$-\sum_{n=1}^{N} \sum_{k=1}^{K} 1_{y_n=k} \mathbf{w}_k \mathbf{x}_n + \sum_{n=1}^{N} \log \sum_{k=1}^{K} \exp(\mathbf{w}_k^\top \mathbf{x}_n).$$

We have already shown that a sum of convex functions is convex, so we only need to show that the following is convex.

$$-\sum_{k=1}^{K} 1_{y_n=k} \mathbf{w}_k \mathbf{x}_n + \log \sum_{k=1}^{K} \exp(\mathbf{w}_k^\top \mathbf{x}_n).$$

The first part is a linear function, which is convex. We only need to prove that the following is convex.

$$\log \sum_{k=1}^{K} \exp(\mathbf{w}_k^\top \mathbf{x}_n)$$

This form is know as a log-sum-exp, and you may know that it is convex. It would be perfectly fine to use this as a fact, but we will prove it using the definition of convexity for the sake of completeness.

To prove: We want to show that for all sets of weights $\mathbf{A} = \mathbf{a}_1, ..., \mathbf{a}_k, \mathbf{B} = \mathbf{b}_1, ..., \mathbf{b}_k$, we have that

$$\lambda \log \left(\sum_k e^{\mathbf{a}_k^\top \mathbf{x}} \right) + (1-\lambda) \log \left(\sum_k e^{\mathbf{b}_k^\top \mathbf{x}} \right) \geq \log \left(\sum_k e^{\lambda \mathbf{a}_k^\top \mathbf{x}} e^{(1-\lambda)\mathbf{b}_k^\top \mathbf{x}} \right).$$

Simplifying the expression: First, we use the following properties of the $\log$, $y \log x = \log x^y$ and $\log x + \log y = \log xy$, to get to the following expression

$$\log \left(\left(\sum_k e^{\mathbf{a}_k^\top \mathbf{x}} \right)^\lambda \left(\sum_k e^{\mathbf{b}_k^\top \mathbf{x}} \right)^{(1-\lambda)} \right) \geq \log \left(\sum_k e^{\lambda \mathbf{a}_k^\top \mathbf{x}} e^{(1-\lambda)\mathbf{b}_k^\top \mathbf{x}} \right).$$

We will now prove this

$$\left(\sum_k e^{\mathbf{a}_k^\top \mathbf{x}} \right)^\lambda \left(\sum_k e^{\mathbf{b}_k^\top \mathbf{x}} \right)^{(1-\lambda)} \geq \sum_k e^{\lambda \mathbf{a}_k^\top \mathbf{x}} e^{(1-\lambda)\mathbf{b}_k^\top \mathbf{x}}.$$

Notice that in $\left(\sum_k e^{\mathbf{a}_k^\top \mathbf{x}} \right)^\lambda$, we are summing over positive numbers due to the exponential. In general, we have that $(\sum_i x_i)^y \geq \sum_i x_i^y$ if all the x_i and y are non negative. Applying this to the left hand side, we have

$$\left(\sum_k e^{\mathbf{a}_k^\top \mathbf{x}} \right)^\lambda \left(\sum_k e^{\mathbf{b}_k^\top \mathbf{x}} \right)^{(1-\lambda)} \geq \left(\sum_k e^{\lambda \mathbf{a}_k^\top \mathbf{x}} \right) \left(\sum_k e^{(1-\lambda)\mathbf{b}_k^\top \mathbf{x}} \right)$$

Now, we will rewrite the sum by apply the following transformation: $(\sum_i x_i)(\sum_i y_i) = \sum_i x_i y_i + \sum_i \sum_{j \neq i} x_i y_j$. This gets us

$$\left(\sum_k e^{\mathbf{a}_k^\top \mathbf{x}} \right)^\lambda \left(\sum_k e^{\mathbf{b}_k^\top \mathbf{x}} \right)^{(1-\lambda)} \geq \sum_k e^{\lambda \mathbf{a}_k^\top \mathbf{x}} e^{(1-\lambda)\mathbf{b}_k^\top \mathbf{x}} + \sum_i \sum_{j \neq i} e^{\lambda \mathbf{a}_i^\top \mathbf{x}} e^{(1-\lambda)\mathbf{b}_j^\top \mathbf{x}}$$

Notice that in the last term, we are summing over non negative numbers, so it is at least 0 and we have another lower bound,

$$\left(\sum_k e^{\mathbf{a}_k^\top \mathbf{x}} \right)^\lambda \left(\sum_k e^{\mathbf{b}_k^\top \mathbf{x}} \right)^{(1-\lambda)} \geq \sum_k e^{\lambda \mathbf{a}_k^\top \mathbf{x}} e^{(1-\lambda)\mathbf{b}_k^\top \mathbf{x}} + \sum_i \sum_{j \neq i} e^{\lambda \mathbf{a}_i^\top \mathbf{x}} e^{(1-\lambda)\mathbf{b}_j^\top \mathbf{x}} \geq \sum_k e^{\lambda \mathbf{a}_k^\top \mathbf{x}} e^{(1-\lambda)\mathbf{b}_k^\top \mathbf{x}}$$

Which concludes the proof