Labs
Machine Learning Course
Fall 2019

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

Problem Set 6, Oct 24, 2019 (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. Using convexity of f, we know that

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

Further since g is increasing, we can apply g on both sides of the above equation to get

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

Finally, using the convexity of g we get

$$\begin{aligned} g(f(\theta x + (1-\theta)y)) &\leq g(\theta f(x) + (1-\theta) f(y)) \\ &\leq \theta g(f(x)) + (1-\theta) g(f(y))\,. \end{aligned}$$

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. Also we can check the convexity by second derivative: for a twice differentiable function of a single variable, if the second derivative is greater than or equal to zero for its entire domain, then the function is convex.

6. 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 means neither points $x^\star$ and $y^\star$ are global minima. This contradicts the initial assumption and proves that a strictly convex function has a unique global minimizer.

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 define $\mathbf{u}_k = e^{\mathbf{a}_k^\top \mathbf{x}}$ and $\mathbf{v}_k = e^{\mathbf{b}_k^\top \mathbf{x}}$, where $\mathbf{u}_k > 0$ and $\mathbf{v}_k > 0$. Thus,

$$\log \left(\sum_k e^{\lambda \mathbf{a}_k^\top \mathbf{x}} e^{(1-\lambda) \mathbf{b}_k^\top \mathbf{x}} \right) = \log \left(\sum_k \left(e^{\mathbf{a}_k^\top \mathbf{x}} \right)^\lambda \left(e^{\mathbf{b}_k^\top \mathbf{x}} \right)^{1-\lambda} \right) = \log \left(\sum_k (\mathbf{u}_k)^\lambda (\mathbf{v}_k)^{1-\lambda} \right),$$

and we would like to prove

$$\lambda \log \left(\sum_k \mathbf{u}_k \right) + (1-\lambda) \log \left(\sum_k \mathbf{v}_k \right) \geq \log \left(\sum_k \mathbf{u}_k^\lambda \mathbf{v}_k^{1-\lambda} \right).$$

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

We can apply this inequality to $\log \left(\sum_k \mathbf{u}_k^\lambda \mathbf{v}_k^{1-\lambda} \right)$, i.e.,

$$\log \left(\sum_k \mathbf{u}_k^\lambda \mathbf{v}_k^{1-\lambda} \right) = \log \left(\sum_k |\mathbf{u}_k^\lambda| |\mathbf{v}_k^{1-\lambda}| \right) \leq \log \left(\left(\sum_k |\mathbf{u}_k^\lambda|^{\frac{1}{\lambda}} \right)^\lambda \left(\sum_k |\mathbf{v}_k^{1-\lambda}|^{\frac{1}{1-\lambda}} \right)^{1-\lambda} \right),$$

where the right formula can be reduced to:

$$\log \left(\left(\sum_k \mathbf{u}_k \right)^\lambda \left(\sum_k \mathbf{v}_k \right)^{1-\lambda} \right) = \lambda \log \left(\sum_k \mathbf{u}_k \right) + (1-\lambda) \log \left(\sum_k \mathbf{v}_k \right).$$

As a result,

$$\log \left(\sum_k \mathbf{u}_k^\lambda \mathbf{v}_k^{1-\lambda} \right) \leq \lambda \log \left(\sum_k \mathbf{u}_k \right) + (1-\lambda) \log \left(\sum_k \mathbf{v}_k \right),$$

which concludes the proof.

3 Mixture of Linear Regression

1. Likelohood: $p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{r}_n) = \prod_{k=1}^{K}[\mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)]^{r_{nk}}$.
2. Joint likelihood: $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{r}) = \prod_{n=1}^{N}\prod_{k=1}^{K}[\mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)]^{r_{nk}}$.
3. Write the joint, then the conditional, and plug in.

$$\begin{aligned}
p(y_n|\boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{\pi}) &= \sum_{k=1}^{K} p(y_n, r_n = k|\boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{\pi}) = \sum_{k=1}^{K} p(y_n|r_n = k, \boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{\pi}) p(r_n = k|\boldsymbol{\pi}) \\
&= \sum_{k=1}^{K} p(y_n|r_n = k, \boldsymbol{x}_n, \boldsymbol{w}, \boldsymbol{\pi})\pi_k = \sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)\pi_k
\end{aligned}$$

4.

$$\begin{aligned}
-\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{\pi}) &= -\log \prod_{n=1}^{N}\sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)\pi_k \\
&= -\sum_{n=1}^{N} \log \sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)\pi_k
\end{aligned}$$

5. (a) A model is identifiable iff $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \rightarrow P_{\boldsymbol{\theta}_1} = P_{\boldsymbol{\theta}_2}$, i.e., the relationship of the parameters to the model is one to one. The given model is not identifiable for two reasons:
   - By permutation of labels.
   - Imagine two models with $\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}_n} = 0 \quad \forall (k, n)$, which are identical except for two different sets of $\boldsymbol{\pi}$: $\boldsymbol{\pi}^*$ and $\hat{\boldsymbol{\pi}}$.

$$\begin{aligned}
\text{Then} \quad & \sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)\pi_k^* = \sum_{k=1}^{K} \mathcal{N}(y_n|\boldsymbol{w}_k^\top \tilde{\boldsymbol{x}}_n, \sigma^2)\hat{\pi}_k \,, \\
\text{implies} \quad & \sum_{k=1}^{K} \mathcal{N}(y_n|0, \sigma^2)\pi_k^* = \sum_{k=1}^{K} \mathcal{N}(y_n|0, \sigma^2)\hat{\pi}_k \,, \\
\text{implies} \quad & \mathcal{N}(y_n|0, \sigma^2)\sum_{k=1}^{K} \pi_k^* = \mathcal{N}(y_n|0, \sigma^2)\sum_{k=1}^{K} \hat{\pi}_k \,, \\
\text{implies} \quad & \mathcal{N}(y_n|0, \sigma^2) = \mathcal{N}(y_n|0, \sigma^2) \,,
\end{aligned}$$

   where the last statement is of course true. Since we started with a set of different $\boldsymbol{\pi}$, this should *never* happen if the model was identifiable.

   (b) The model is also not *convex*, since a sum of Gaussians is not convex.