

Problem Set 6, Oct 20, 2020 (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

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

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 x and y be two elements in the domain. Let $x = \mathbf{w}^\top \mathbf{x} + b$ and $y = \mathbf{w}^\top \mathbf{y} + b$. 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^* and y^* . Let $z^* = (x^* + y^*)/2$. Then, since f is strictly convex, we have $f(z^*) < \frac{1}{2}(f(x^*) + f(y^*)) = f(x^*) = f(y^*)$, which means neither points x^* and y^* 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 $\mathbf{W} = \mathbf{w}_1, \dots, \mathbf{w}_K$ 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.

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

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.

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

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

2. We get

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

Where $\text{softmax}(n, k) = \frac{\exp(\eta_{nk})}{\sum_{j=1}^K \exp(\eta_{nj})}$.

3. The negative of the log-likelihood is

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

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

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

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 known 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 $\mathbf{A} = \mathbf{a}_1, \dots, \mathbf{a}_K, \mathbf{B} = \mathbf{b}_1, \dots, \mathbf{b}_K$, we have that

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

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,

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

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

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

4.

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

5. (a) The model is not *convex* in general. E.g., consider the case when $N = 1$, $K = 2$. Then negative log-likelihood is equal to

$$\frac{1}{2} \log 2\pi\sigma^2 - \log \left[\exp \left(-\frac{(y - \mathbf{w}_1^\top \mathbf{x})^2}{2\sigma^2} \right) \pi_1 + \exp \left(-\frac{(y - \mathbf{w}_2^\top \mathbf{x})^2}{2\sigma^2} \right) (1 - \pi_1) \right]$$

The first term is a constant, we will look only at the second term and prove that it is not convex. Define

$$f(\mathbf{w}_1, \mathbf{w}_2, \pi_1) := -\log \left[\exp \left(-\frac{(y - \mathbf{w}_1^\top \mathbf{x})^2}{2\sigma^2} \right) \pi_1 + \exp \left(-\frac{(y - \mathbf{w}_2^\top \mathbf{x})^2}{2\sigma^2} \right) (1 - \pi_1) \right]$$

In order to prove that $f(\mathbf{w}_1, \mathbf{w}_2, \pi_1)$ is not convex we will construct two points $p^1 = (\mathbf{w}_1^1, \mathbf{w}_2^1, \pi_1^1)$ and $p^2 = (\mathbf{w}_1^2, \mathbf{w}_2^2, \pi_1^2)$ such that $f(\frac{1}{2}p^1 + \frac{1}{2}p^2) > \frac{1}{2}f(p^1) + \frac{1}{2}f(p^2)$.

Let

$$p^1 = \left(\frac{y}{\|\mathbf{x}\|_2^2} \mathbf{x}, \frac{y+2}{\|\mathbf{x}\|_2^2} \mathbf{x}, 1 \right) \quad p^2 = \left(\frac{y+2}{\|\mathbf{x}\|_2^2} \mathbf{x}, \frac{y}{\|\mathbf{x}\|_2^2} \mathbf{x}, 0 \right),$$

note that $\mathbf{x} \neq \mathbf{0}$ since its first coordinate is equal to 1 as stated in the exercise. Then

$$\begin{aligned} f(p^1) &= -\log \left[\exp \left(-\frac{0}{2\sigma^2} \right) \right] = 0 \\ f(p^2) &= -\log \left[\exp \left(-\frac{0}{2\sigma^2} \right) \right] = 0 \\ f\left(\frac{1}{2}p^1 + \frac{1}{2}p^2\right) &= -\log \left[\exp \left(-\frac{1}{2\sigma^2} \right) \right] = \frac{1}{2\sigma^2} > 0 \end{aligned}$$

This proves that negative log-likelihood is not convex in general.

- (b) The given model is not identifiable by permutation of indexes of mixture components.

Assume that the model is identifiable and true solution is \mathbf{w}^* , $\boldsymbol{\pi}^*$ is found by MLE when the data size grows to infinity, i.e.

$$\mathbf{w}^*, \boldsymbol{\pi}^* = \arg \min_{\mathbf{w}, \boldsymbol{\pi}} [L(\mathbf{w}, \boldsymbol{\pi}) := -\log p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \boldsymbol{\pi})]$$

Then we will construct the second point $\hat{\mathbf{w}}, \hat{\boldsymbol{\pi}} \neq \mathbf{w}^*, \boldsymbol{\pi}^*$ such that $L(\hat{\mathbf{w}}, \hat{\boldsymbol{\pi}}) = L(\mathbf{w}^*, \boldsymbol{\pi}^*)$. This would mean that $\hat{\mathbf{w}}, \hat{\boldsymbol{\pi}}$ is also a solution of MLE and there is no way to distinguish between the true solution $\mathbf{w}^*, \boldsymbol{\pi}^*$ and a point $\hat{\mathbf{w}}, \hat{\boldsymbol{\pi}}$, so MLE doesn't always give a true solution.

We define $\hat{\mathbf{w}}, \hat{\boldsymbol{\pi}}$ as follows

$$\begin{aligned}\hat{\mathbf{w}}_1 &= \mathbf{w}_2^* & \hat{\boldsymbol{\pi}}_1 &= \boldsymbol{\pi}_2^* \\ \hat{\mathbf{w}}_2 &= \mathbf{w}_1^* & \hat{\boldsymbol{\pi}}_2 &= \boldsymbol{\pi}_1^* \\ \hat{\mathbf{w}}_i &= \mathbf{w}_i^*, i \geq 3 & \hat{\boldsymbol{\pi}}_i &= \boldsymbol{\pi}_i^*, i \geq 3,\end{aligned}$$

i.e. vectors corresponding to the first two mixture components are permuted.

Then indeed the losses at these two points are equal,

$$L(\hat{\mathbf{w}}, \hat{\boldsymbol{\pi}}) = - \sum_{n=1}^N \log \sum_{k=1}^K \mathcal{N}(y_n | \hat{\mathbf{w}}_k^\top \tilde{\mathbf{x}}_n, \sigma^2) \hat{\pi}_k = - \sum_{n=1}^N \log \sum_{k=1}^K \mathcal{N}(y_n | \mathbf{w}_k^{*\top} \tilde{\mathbf{x}}_n, \sigma^2) \pi_k^* = L(\mathbf{w}^*, \boldsymbol{\pi}^*).$$

This ends the proof.