COMS 4721 Spring 2020 Homework 3 Written Assignment: Optimization, Logistic Regression, SVM

Instruction

Please prepare your write-up as a typeset PDF document (which can be generated using LaTex or Word). If you choose to hand-write certain portions of the assignment (**all source code must be typeset**), make sure your handwriting is legible and the consistency in the format with other pages which are typeset (e.g. page size, page numbering). If we cannot read your handwriting, you may not receive credit for the question.

This write-up contains all of your supporting materials which include plots, source code, and proofs for each of the parts in the assignment. Submit your assignment on Gradescope by clearly marking the pages for each part. On the first page of your write-up, please typeset: (1) your name and your UNI; (2) all of your collaborators whom you discussed the assignment with; (3) the parts of the assignment you had collaborated on. The solutions to the problems need to start from the second page. Please write up the solutions by yourself. The academic rules of conduct is found in the course syllabus.

Suggestions

If necessary, please define notations and explain reasoning behind the solutions as concisely as possible. Solutions without explanations when needed may receive no credit. Points can be deducted for solutions with unnecessarily long explanations for lack of clarity. Source code comment can be useful for explaining the logic behind your solutions. Please start early!

Problem 1 (30 points)

More logistic regression: In this problem we will revisit our old friend logistic regression:

(a) (10 points) Recall that the logistic function is given by:

$$g(z) = \frac{1}{1 + \exp(-z)}$$

Prove that $g(-z) = 1 - g(z)$ and its inverse is given by: $g^{-1}(y) = \ln \frac{y}{1-y}$. Also prove that $\frac{dg(z)}{dz} = g(z)(1 - g(z))$.

Proof.

$$\begin{aligned} g(-z) &= \frac{1}{1 + \exp(z)} \\ &= 1 - \frac{\exp(z)}{1 + \exp(z)} \\ &= 1 - \frac{1}{1 + \exp(-z)} \\ &= 1 - g(z) \end{aligned}$$

To find its inverse:

$$\begin{aligned} y &= \frac{1}{1 + \exp(-z)} \\ \exp(-z) &= \frac{1}{y} - 1 \\ -z &= \ln \frac{1 - y}{y} \\ z &= \ln \frac{y}{1 - y} \end{aligned}$$

Therefore $g^{-1}(y) = \ln \frac{y}{1-y}$.

To prove that $\frac{dg(z)}{dz} = g(z)(1 - g(z))$:

$$\begin{aligned} \frac{dg(z)}{dz} &= \frac{d\frac{1}{1+\exp(-z)}}{dz} \\ &= \frac{0 - \exp(-z)(-1)}{(1 + \exp(-z))^2} \\ &= \frac{1}{1 + \exp(-z)} \frac{\exp(-z)}{1 + \exp(-z)} \\ &= g(z)(1 - g(z)) \end{aligned}$$

□

(b) (10 points) Here is a way to extend logistic regression to handle more than two labels, i.e. learn $\hat{f} : \mathcal{X} \to \mathcal{Y}$ where $\mathcal{Y} = \{0, 1, \cdots, K-1\}$ and K is the number of categories. Multinomial logit model fits $K-1$ linear models:

$$\ln \frac{\Pr(y=0 \mid \mathbf{x})}{\Pr(y=K-1 \mid \mathbf{x})} = \mathbf{w_0}^T \mathbf{x}$$
$$\vdots$$
$$\ln \frac{\Pr(y=K-2 \mid \mathbf{x})}{\Pr(y=K-1 \mid \mathbf{x})} = \mathbf{w_{K-2}}^T \mathbf{x}$$

by using the last label $K-1$ as the pivot. Show that we can define:

$$\Pr(y=0 \mid \mathbf{x}) = \frac{\exp(\mathbf{w_0}^T \mathbf{x})}{1 + \sum_{k=0}^{K-2} \exp(\mathbf{w_k}^T \mathbf{x})}$$
$$\vdots = \vdots$$
$$\Pr(y=K-2 \mid \mathbf{x}) = \frac{\exp(\mathbf{w_{K-2}}^T \mathbf{x})}{1 + \sum_{k=0}^{K-2} \exp(\mathbf{w_k}^T \mathbf{x})}$$
$$\Pr(y=K-1 \mid \mathbf{x}) = \frac{1}{1 + \sum_{k=0}^{K-2} \exp(\mathbf{w_k}^T \mathbf{x})}$$

Hint: Use the fact that $\sum_{k=0}^{K-1} \Pr(y=k \mid \mathbf{x}) = 1$.

Proof. We are given that for $i = 0, 1, 2 \ldots K-2$

$$\ln \frac{\Pr(y=i \mid \mathbf{x})}{\Pr(y=K-1 \mid \mathbf{x})} = \mathbf{w_i}^T \mathbf{x}$$
$$\frac{\Pr(y=i \mid \mathbf{x})}{\Pr(y=K-1 \mid \mathbf{x})} = \exp(\mathbf{w_i}^T \mathbf{x})$$
$$\Pr(y=i \mid \mathbf{x}) = \Pr(y=K-1 \mid \mathbf{x}) \exp(\mathbf{w_i}^T \mathbf{x})$$

From the property of probability, we also know that

$$\sum_{i=0}^{K-1} \Pr(y=i \mid \mathbf{x}) = 1$$

Therefore

$$\sum_{i=0}^{K-2} \Pr(y = K-1 \mid \mathbf{x}) \exp(\mathbf{w_i}^T \mathbf{x}) + \Pr(y = K-1 \mid \mathbf{x}) = 1$$

$$\Pr(y = K-1 \mid \mathbf{x}) \left[1 + \sum_{i=0}^{K-2} \exp(\mathbf{w_i}^T \mathbf{x}) \right] = 1$$

$$\Pr(y = K-1 \mid \mathbf{x}) = \frac{1}{1 + \sum_{i=0}^{K-2} \exp(\mathbf{w_i}^T \mathbf{x})}$$

Since for $i = 0, 1, 2 \ldots K-2$ we have

$$\Pr(y = i \mid \mathbf{x}) = \Pr(y = K-1 \mid \mathbf{x}) \exp(\mathbf{w_i}^T \mathbf{x})$$

We obtain for $i = 0, 1, 2 \ldots K-2$

$$\Pr(y = i \mid \mathbf{x}) = \frac{\exp(\mathbf{w_i}^T \mathbf{x})}{1 + \sum_{k=0}^{K-2} \exp(\mathbf{w_k}^T \mathbf{x})}$$

And for $i = K-1$:

$$\Pr(y = K-1 \mid \mathbf{x}) = \frac{1}{1 + \sum_{i=0}^{K-2} \exp(\mathbf{w_i}^T \mathbf{x})}$$

To summarize:

$$\Pr(y = 0 \mid \mathbf{x}) = \frac{\exp(\mathbf{w_0}^T \mathbf{x})}{1 + \sum_{k=0}^{K-2} \exp(\mathbf{w_k}^T \mathbf{x})}$$

$$\vdots = \vdots$$

$$\Pr(y = K-2 \mid \mathbf{x}) = \frac{\exp(\mathbf{w_{K-2}}^T \mathbf{x})}{1 + \sum_{k=0}^{K-2} \exp(\mathbf{w_k}^T \mathbf{x})}$$

$$\Pr(y = K-1 \mid \mathbf{x}) = \frac{1}{1 + \sum_{k=0}^{K-2} \exp(\mathbf{w_k}^T \mathbf{x})}$$

□

(c) (10 points) For $0 \leq k_1 < k_2 \leq K-1$, show that for multinomial logit, the decision boundary between the class label k_1 and the class label k_2 is linear.

Proof. The decision boundary for multinomial logit is

$$\ln \frac{\Pr(y = k_1 \mid \mathbf{x})}{\Pr(y = k_2 \mid \mathbf{x})}$$

For $k_2 = K - 1$, we already have the default K-1 models have log odd ratio between the class k1 against the class Kâ1 for $0 <= k_1 <= K - 2$:

$$\ln \frac{\Pr(y = 0 \mid \mathbf{x})}{\Pr(y = K - 1 \mid \mathbf{x})} = \mathbf{w_0}^T \mathbf{x}$$

$$\vdots$$

$$\ln \frac{\Pr(y = K - 2 \mid \mathbf{x})}{\Pr(y = K - 1 \mid \mathbf{x})} = \mathbf{w_{K-2}}^T \mathbf{x}$$

All of the models are linear.
For $k_2 \neq K - 1$, we have:

$$\begin{aligned}
\ln \frac{\Pr(y = k_1 \mid \mathbf{x})}{\Pr(y = k_2 \mid \mathbf{x})} &= \ln \frac{\frac{\exp(\mathbf{w_{k_1}}^T \mathbf{x})}{1 + \sum_{k=0}^{K-2} \exp(\mathbf{w_k}^T \mathbf{x})}}{\frac{\exp(\mathbf{w_{k_2}}^T \mathbf{x})}{1 + \sum_{k=0}^{K-2} \exp(\mathbf{w_k}^T \mathbf{x})}} \\
&= \ln \frac{\exp(\mathbf{w_{k_1}}^T \mathbf{x})}{\exp(\mathbf{w_{k_2}}^T \mathbf{x})} \\
&= \ln \exp(\mathbf{w_{k_1}}^T - \mathbf{w_{k_2}}^T) \mathbf{x} \\
&= (\mathbf{w_{k_1}}^T - \mathbf{w_{k_2}}^T) \mathbf{x}
\end{aligned}$$

The result is a linear equation. Therefore we can conclude that for $0 \leq k_1 < k_2 \leq K - 1$ and for multinomial logit, the decision boundary between the class label k_1 and the class label k_2 is linear. □

Problem 2 (30 points)

Optimization: Some problems on optimization.

(a) (10 points) Prove Jensen's inequality: For any $x_1, \cdots, x_N > 0$ and $\lambda_1, \cdots, \lambda_N \geq 0$ with $\sum_{i=1}^{N} \lambda_i = 1$, For a convex function $f : \mathbb{R} \to \mathbb{R}$,

$$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \leq \sum_{i=1}^{N} \lambda_i f(x_i)$$

For a concave function $f : \mathbb{R} \to \mathbb{R}$,

$$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \geq \sum_{i=1}^{N} \lambda_i f(x_i)$$

Hint: For $N = 2$, use the definition of convexity/concavity. For $N > 2$, use induction.
Answer:

Proof. Convex case: For $N = 2$, we know in lecture that for a convex function $f : \mathbb{R} \to \mathbb{R}$, f is below any line segment between two points on f; that is, for any $x_1, x_2 > 0$, $\forall t \in [0, 1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

Let $\lambda_1 = t, \lambda_2 = 1 - t$
We can re-write the inequality above:
For any $x_1, x_2 > 0$ and $\lambda_1, \lambda_2 \geq 0$ with $\sum_{i=1}^{2} \lambda_i = 1$, for a convex function $f : \mathbb{R} \to \mathbb{R}$,

$$f\left(\sum_{i=1}^{2} \lambda_i x_i\right) \leq \sum_{i=1}^{2} \lambda_i f(x_i)$$

For $N > 2$, we can use induction by assuming the inequality holds for $N = k$, that is:
For any $x_1, \cdots, x_k > 0$ and $\lambda_1, \cdots, \lambda_k \geq 0$ with $\sum_{i=1}^{k} \lambda_i = 1$, for a convex function $f : \mathbb{R} \to \mathbb{R}$,

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \leq \sum_{i=1}^{k} \lambda_i f(x_i)$$

For any $x_1, \cdots, x_{k+1} > 0$ and $\lambda_1, \cdots, \lambda_k \geq 0$ with $\sum_{i=1}^{k+1} \lambda_i = 1$, we have

$$f\left(\sum_{i=1}^{k+1} \lambda_i x_i\right) = f\left(\lambda_1 x_1 + \sum_{i=2}^{k} \lambda_i x_i\right) = f\left(\lambda_1 x_1 + (1-\lambda_1)\frac{\sum_{i=2}^{k} \lambda_i x_i}{1-\lambda_1}\right) \leq \sum_{i=1}^{k} \lambda_i f(x_i)$$

By the assumption, we have

$$f\left(\lambda_1 x_1 + (1-\lambda_1)\frac{\sum_{i=2}^{k}\lambda_i x_i}{1-\lambda_1}\right) \leq \lambda_1 f(x_1) + (1-\lambda_1) f\left(\frac{\sum_{i=2}^{k}\lambda_i}{1-\lambda_1} x_i\right)$$

Since $\frac{\sum_{i=2}^{\lambda_i}}{1-\lambda_i} = 1$, we can again employ the assumption and obtain:

$$f\left(\lambda_2 x_2 + (1-\lambda_2)\frac{\sum_{i=3}^{k}\lambda_i x_i}{1-\lambda_2}\right) \leq \lambda_2 f(x_2) + (1-\lambda_2) f\left(\frac{\sum_{i=3}^{k}\lambda_i}{1-\lambda_3} x_i\right)$$

Again we can apply the assumption and obtain:

$$f\left(\lambda_3 x_3 + (1-\lambda_3)\frac{\sum_{i=3}^{k}\lambda_i x_i}{1-\lambda_2}\right) \leq \lambda_2 f(x_2) + (1-\lambda_2) f\left(\frac{\sum_{i=3}^{k}\lambda_i}{1-\lambda_3} x_i\right)$$

$$\ldots$$

Eventually we obtain

$$f\left(\lambda_1 x_1 + (1-\lambda_1)\frac{\sum_{i=2}^{k}\lambda_i x_i}{1-\lambda_1}\right) \leq \lambda_1 f(x_1) + \ldots + \lambda_{k+1} f(x_{k+1})$$

In other words,

$$f\left(\sum_{i=1}^{N}\lambda_i x_i\right) \leq \sum_{i=1}^{N}\lambda_i f(x_i) \quad \square$$

Proof. Concave case: For $N = 2$, we know in lecture that for a concave function $f : \mathbb{R} \to \mathbb{R}$, f is above any line segment between two points on f; that is, for any $x_1, x_2 > 0$, $\forall t \in [0, 1]$

$$f(tx_1 + (1-t)x_2) \geq tf(x_1) + (1-t)f(x_2)$$

Let $\lambda_1 = t, \lambda_2 = 1 - t$
We can re-write the inequality above:
For any $x_1, x_2 > 0$ and $\lambda_1, \lambda_2 \geq 0$ with $\sum_{i=1}^{2}\lambda_i = 1$, For a concave function $f : \mathbb{R} \to \mathbb{R}$,

$$f\left(\sum_{i=1}^{2}\lambda_i x_i\right) \geq \sum_{i=1}^{2}\lambda_i f(x_i)$$

For $N > 2$, we can use induction by assuming the inequality holds for $N = k$, that is:
For any $x_1, \cdots, x_k > 0$ and $\lambda_1, \cdots, \lambda_k \geq 0$ with $\sum_{i=1}^{k} \lambda_i = 1$, for a concave function $f : \mathbb{R} \to \mathbb{R}$,

$$f\left(\sum_{i=1}^{k} \lambda_i x_i\right) \geq \sum_{i=1}^{k} \lambda_i f(x_i)$$

For any $x_1, \cdots, x_{k+1} > 0$ and $\lambda_1, \cdots, \lambda_k \geq 0$ with $\sum_{i=1}^{k+1} \lambda_i = 1$, we have

$$f\left(\sum_{i=1}^{k+1} \lambda_i x_i\right) = f\left(\lambda_1 x_1 + \sum_{i=2}^{k} \lambda_i x_i\right) = f\left(\lambda_1 x_1 + (1-\lambda_1)\frac{\sum_{i=2}^{k} \lambda_i x_i}{1-\lambda_1}\right) \geq \sum_{i=1}^{k} \lambda_i f(x_i)$$

By the assumption, we have

$$f\left(\lambda_1 x_1 + (1-\lambda_1)\frac{\sum_{i=2}^{k} \lambda_i x_i}{1-\lambda_1}\right) \geq \lambda_1 f(x_1) + (1-\lambda_1) f\left(\frac{\sum_{i=2}^{k} \lambda_i}{1-\lambda_1} x_i\right)$$

Since $\frac{\sum_{i=2}^{k} \lambda_i}{1-\lambda_i} = 1$, we can again employ the assumption and obtain:

$$f\left(\lambda_2 x_2 + (1-\lambda_2)\frac{\sum_{i=3}^{k} \lambda_i x_i}{1-\lambda_2}\right) \geq \lambda_2 f(x_2) + (1-\lambda_2) f\left(\frac{\sum_{i=3}^{k} \lambda_i}{1-\lambda_3} x_i\right)$$

Again we can apply the assumption and obtain:

$$f\left(\lambda_3 x_3 + (1-\lambda_3)\frac{\sum_{i=3}^{k} \lambda_i x_i}{1-\lambda_2}\right) \geq \lambda_2 f(x_2) + (1-\lambda_2) f\left(\frac{\sum_{i=3}^{k} \lambda_i}{1-\lambda_3} x_i\right)$$

$$\cdots$$

Eventually we obtain

$$f\left(\lambda_1 x_1 + (1-\lambda_1)\frac{\sum_{i=2}^{k} \lambda_i x_i}{1-\lambda_1}\right) \geq \lambda_1 f(x_1) + \ldots + \lambda_{k+1} f(x_{k+1})$$

In other words,

$$f\left(\sum_{i=1}^{N} \lambda_i x_i\right) \geq \sum_{i=1}^{N} \lambda_i f(x_i) \quad \square$$

(b) (10 points) Prove that for any N positive real numbers, the arithmetic mean is at least the geometric mean. That is, for any $x_1, \cdots, x_N > 0$ and $\lambda_1, \cdots, \lambda_N \geq 0$ with $\sum_{i=1}^{N} \lambda_i = 1$:

$$\sum_{i=1}^{N} \lambda_i x_i \geq \prod_{i=1}^{N} x_i^{\lambda_i}$$

Hint: Use Jensen's inequality.

Proof. From second order condition for concavity, we can show that $f(x) = \ln(x)$ is strictly concave for for yany $x_1, \cdots, x_N > 0$:

$$f''(x) = -\frac{1}{x^2} < 0$$

Using Jensen's inequality, for any $x_1, \cdots, x_N > 0$ and $\lambda_1, \cdots, \lambda_N \geq 0$ with $\sum_{i=1}^{N} \lambda_i = 1$ we obtain:

$$\ln\left(\sum_{i=1}^{N} \lambda_i x_i\right) \geq \sum_{i=1}^{N} \lambda_i \ln(x_i)$$

$$\exp(\ln\left(\sum_{i=1}^{N} \lambda_i x_i\right)) \geq \exp(\sum_{i=1}^{N} \lambda_i \ln(x_i))$$

$$\left(\sum_{i=1}^{N} \lambda_i x_i\right) \geq \exp(\sum_{i=1}^{N} \ln(x_i^{\lambda_i}))$$

$$\left(\sum_{i=1}^{N} \lambda_i x_i\right) \geq \exp(\ln(\prod_{i=1}^{N} x_i^{\lambda_i}))$$

$$\sum_{i=1}^{N} \lambda_i x_i \geq \prod_{i=1}^{N} x_i^{\lambda_i}$$

□

(c) (10 points) Prove that for $\theta, f_i \in \mathbb{R}^d$:

$$\sum_{i=1}^{K} \exp(\theta^T f_i) \geq \exp\left(\sum_{i=1}^{K} \lambda_i \theta^T f_i - \sum_{i=1}^{K} \lambda_i \ln \lambda_i\right)$$

where $\lambda_i = \frac{\exp(\theta^T f_i)}{\sum j=1^K \exp(\theta^T f_j)}$.

Hint: Express $\sum_{i=1}^{K} \exp(\theta^T f_i) = \sum_{i=1}^{K} \lambda_i \left(\frac{\exp(\theta^T f_i)}{\lambda_i}\right)$ and note $\sum_{i=1}^{K} \lambda_i = 1$. Choose $f(x) = \ln x$ and invoke the concave version of the Jensen's inequality on $\ln\left(\sum_{i=1}^{K} \lambda_i \left(\frac{\exp(\theta^T f_i)}{\lambda_i}\right)\right)$

Proof. Using Jensen's inequality, for any $x_1, \cdots, x_N > 0$ and $\lambda_1, \cdots, \lambda_K \geq 0$ with $\sum_{i=1}^{K} \lambda_i = 1$ we obtain:

$$\ln\left(\sum_{i=1}^{K} \lambda_i \left(\frac{\exp(\theta^T f_i)}{\lambda_i}\right)\right) \geq \sum_{i=1}^{K} \lambda_i \ln(\frac{\exp(\theta^T f_i)}{\lambda_i})$$

$$\sum_{i=1}^{K} \lambda_i \left(\frac{\exp(\theta^T f_i)}{\lambda_i}\right) \geq \exp(\sum_{i=1}^{K} \lambda_i \ln(\frac{\exp(\theta^T f_i)}{\lambda_i}))$$

$$\sum_{i=1}^{K} \big(\exp(\theta^T f_i)\big) \geq \exp(\ln \prod_{i=1}^{K} \left(\frac{\exp(\theta^T f_i)}{\lambda_i}\right)^{\lambda_i})$$

$$\sum_{i=1}^{K} \big(\exp(\theta^T f_i)\big) \geq \prod_{i=1}^{K} \left(\frac{\exp(\theta^T f_i)}{\lambda_i}\right)^{\lambda_i}$$

$$\sum_{i=1}^{K} \big(\exp(\theta^T f_i)\big) \geq \prod_{i=1}^{K} \big(\exp(\theta^T f_i - \ln \lambda_i)\big)^{\lambda_i}$$

$$\sum_{i=1}^{K} \big(\exp(\theta^T f_i)\big) \geq \exp\left((\theta^T f_i - \ln \lambda_i) \sum_{i=1}^{K} \lambda_i\right)$$

$$\sum_{i=1}^{K} \big(\exp(\theta^T f_i)\big) \geq \exp\left(\sum_{i=1}^{K} \lambda_i \theta^T f_i - \sum_{i=1}^{K} \lambda_i \ln \lambda_i\right)$$

□

Problem 3 (10 points)

Support Vector Machines Suppose we are given the data $\{(\mathbf{x}^{(i)} \in \mathbb{R}^d, y^{(i)}) \in \{-1, 1\}\}_{i=1}^N$. (note that the label set is now $\{-1, 1\}$). Suppose there exist a linear separator $\mathbf{w}$ such that for all points, $y^{(i)} \cdot \mathbf{w}^T\mathbf{x}^{(i)} \geq 1$. We note that the support vectors lie on the hyperplanes $\mathbf{w}^T\mathbf{x} = 1$ and and $\mathbf{w}^T\mathbf{x} = -1$.

Prove that the distance between these two hyperplanes, $\mathbf{w}^T\mathbf{x} = 1$ and $\mathbf{w}^T\mathbf{x} = -1$, is $\frac{2}{||\mathbf{w}||_2}$.

From this optimization formulation, explain why SVM is a max margin classifier.

Proof. The line $x = \lambda\frac{w}{||\mathbf{w}||_2}$ intersects the two hyperplanes at $\lambda_1 = \frac{1}{||\mathbf{w}||_2}$ and $\lambda_2 = \frac{-1}{||\mathbf{w}||_2}$. It follows that the distance between these two hyperplanes, $\mathbf{w}^T\mathbf{x} = 1$ and $\mathbf{w}^T\mathbf{x} = -1$, is $d = \frac{|1-(-1)|}{||\mathbf{w}||_2} = \frac{2}{||\mathbf{w}||_2}$.

In the hard-margin SVM formulation for linearly separable data, the objective is a constrained optimization problem:

$$\underset{y^{(i)} \cdot \mathbf{w}^T\mathbf{x}^{(i)} \geq 1}{\text{argmin}} \frac{\mathbf{w}^T\mathbf{w}}{2}$$

This is equivalent to the objective:

$$\underset{y^{(i)} \cdot \mathbf{w}^T\mathbf{x}^{(i)} \geq 1}{\text{argmax}} \frac{2}{||\mathbf{w}||_2}$$

Note that $\frac{2}{||\mathbf{w}||_2}$ is the margin between the hyperplanes $\mathbf{w}^T\mathbf{x} = 1$ and $\mathbf{w}^T\mathbf{x} = -1$ that classify data into labels $\{-1, 1\}$. That is the objective is a max margin classifier:

$$\text{argmax} \frac{2}{||\mathbf{w}||_2}$$

$$y = 1 \text{ when } \mathbf{w}^T\mathbf{x}^{(i)} \geq 1$$
$$y = -1 \text{ when } \mathbf{w}^T\mathbf{x}^{(i)} \leq -1$$

Therefore SVM is a max margin classifier.

□