Machine Learning Spring 2019 Homework 2

Homework must be submitted electronically on Canvas. Make sure to explain you reasoning or show your derivations. Except for answers that are especially straightforward, you will lose points for unjustified answers, even if they are correct.

Written Problems

1. Multiclass logistic regression has the form

$$p(y|\boldsymbol{x}; W) := \frac{\exp(\boldsymbol{w}_y^\top \boldsymbol{x})}{\sum_{c=1}^{C} \exp(\boldsymbol{w}_c^\top \boldsymbol{x})} , \tag{1}$$

where W is a set of weight vectors $\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_C\}$ for each class.

(a) (5 points) You have a batch of n data points and labels $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$. Write the conditional log-likelihood of the labels given the features and simplify it. Show and explain the steps you take to simplify the expression. You should end up with an expression very similar to Equation (13).

Solution: (Note because we number equations in order, the equation numbers of the formulas given in the programming assignment have changed in this solutions document.)
The data likelihood is the probability of the entire dataset given the parameters

$$\begin{aligned} p(\{y_1, \ldots, y_n\}|\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}) &= \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i; W) \\ &= \prod_{i=1}^{n} \frac{\exp(\boldsymbol{w}_y^\top \boldsymbol{x}_i)}{\sum_{c=1}^{C} \exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)} . \end{aligned} \tag{2}$$

Taking the log of both sides, we get the log-likelihood and expand the product into the sum of log terms

$$\begin{aligned} \log p(\{y_1, \ldots, y_n\}|\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}) &= \log \prod_{i=1}^{n} \frac{\exp(\boldsymbol{w}_y^\top \boldsymbol{x}_i)}{\sum_{c=1}^{C} \exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)} \\ &= \sum_{i=1}^{n} \log \left(\frac{\exp(\boldsymbol{w}_y^\top \boldsymbol{x}_i)}{\sum_{c=1}^{C} \exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)} \right) \\ &= \sum_{i=1}^{n} \log \left(\exp(\boldsymbol{w}_y^\top \boldsymbol{x}_i) \right) - \log \left(\sum_{c=1}^{C} \exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i) \right) \\ &= \sum_{i=1}^{n} \boldsymbol{w}_y^\top \boldsymbol{x}_i - \sum_{i=1}^{n} \log \left(\sum_{c=1}^{C} \exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i) \right) . \end{aligned} \tag{3}$$

This final expression can't be simplified further. The difference between this and Equation (13) is that it is the positive log-likelihood and we have not added a regularizer.

(b) (5 points) Derive the gradient of the logistic regression likelihood with respect to any one of the $\boldsymbol{w}_c$ vectors. It should look very similar to Equation (14) or Equation (15) below (either form is acceptable for full credit). Show and explain the steps you take to derive the gradient.

Solution: First, we rewrite the objective, the *positive* log-likelihood for convenience, replacing the iterator variable over classes with c' to avoid confusion:

$$L(D) = \sum_{i=1}^{n} \boldsymbol{w}_y^\top \boldsymbol{x}_i - \sum_{i=1}^{n} \log \left(\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i) \right) .$$

We can take the gradient with respect to $\boldsymbol{w}_c$ of each summation separately. The first term $\sum_{i=1}^{n} \boldsymbol{w}_y^\top \boldsymbol{x}_i$ has zero-gradient for $\boldsymbol{w}_c$ when $y \neq c$, and otherwise it's just a linear product of the weight vector. Thus the gradient is

$$\nabla_{\boldsymbol{w}_c} \sum_{i=1}^{n} \boldsymbol{w}_y^\top \boldsymbol{x}_i = \sum_{i=1} I(y_i = c)\boldsymbol{x}_i,$$

where I is the indicator function that is 1.0 if its input is true and 0.0 if it is false. This gradient sums together the input $\boldsymbol{x}_i$s that are in class c.

The second term requires applying the chain rule. We can consider each term inside the summation over i separately. In this case, the $\log$ is our outer function, and the inner function is $\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)$, then the chain rule gives us

$$\frac{d\ \log\left(\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)\right)}{d\ \boldsymbol{w}_c} = \frac{d\ \log\left(\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)\right)}{d\ \left(\sum_{'c=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)\right)} \cdot \frac{d\ \left(\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)\right)}{d\ \boldsymbol{w}_c} \tag{4}$$

The derivative of $\log(x)$ is $\frac{1}{x}$, so we can first simplify to

$$\frac{d\ \log\left(\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)\right)}{d\ \left(\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)\right)} \cdot \frac{d\ \left(\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)\right)}{d\ \boldsymbol{w}_c} = \frac{1}{\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} \cdot \frac{d\ \left(\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)\right)}{d\ \boldsymbol{w}_c} \tag{5}$$

Then the gradient for the second part can be simplified using a few steps. First, note that the summation over c' in the numerator only interacts with $\boldsymbol{w}_c$ when $c' = c$, which allows us to eliminate the sum over c' and only consider the case where $c' = c$

$$\frac{d\ \left(\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)\right)}{d\ \boldsymbol{w}_c} = \frac{d\ \exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}{d\ \boldsymbol{w}_c}$$

Next, we can apply the chain rule again, noting that the derivative of $\exp(x) = \exp(x)$

$$\frac{d\ \exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}{d\ \boldsymbol{w}_c} = \exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)\boldsymbol{x}_i.$$

Putting the pieces back together, we have

$$\begin{aligned} \frac{d\ \log\left(\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)\right)}{d\ \boldsymbol{w}_c} &= \frac{1}{\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} \cdot \exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)\boldsymbol{x}_i \\ &= \frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)\boldsymbol{x}_i}{\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} \end{aligned} \tag{6}$$

Then plugging this back into the full gradient expression, combining summations, and pulling out common factors, we get

$$\begin{aligned} \nabla_{\boldsymbol{w}_c} L(D) &= \sum_{i=1} I(y_i = c)\boldsymbol{x}_i - \sum_{i=1}^{n} \frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)\boldsymbol{x}_i}{\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} \\ &= \sum_{i=1} \left(I(y_i = c)\boldsymbol{x}_i - \frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}{\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} \cdot \boldsymbol{x}_i \right) \\ &= \sum_{i=1} \boldsymbol{x}_i \left(I(y_i = c) - \frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}{\sum_{c'=1}^{C} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} \right) \\ &= \sum_{i=1} \boldsymbol{x}_i \left(I(y_i = c) - p(y_i = c|\boldsymbol{x}_i; W) \right), \end{aligned} \tag{7}$$

where in the last step, we used the fact that the expression being subtracted is the defined conditional probability estimate of the logistic regression model. Your solution did not need to be as thorough as this very precise step-by-step derivation for full credit.

(c) (5 points) We often add a regularizer (as we do below for the programming portion) that penalizes the magnitude of the weight vectors. These regularizers "prefer" the weights to be the all-zeros vector. What is the estimated probability of this all-zeros solution? Explain in one to three sentences why this probability is reasonable as the maximally regularized solution.

Solution: When $\boldsymbol{w}_c = \vec{0}$ for all classes, we get uniform probability for each class independently of the input data.

This uniform prediction is a reasonable output for a maximally regularized solution because it represents the maximum amount of uncertainty. Given no other information (e.g., no data or completely random, noisy data) it makes sense to simply predict uniform probabilities, since nothing has told us to do otherwise.

It's possible to make a variety of arguments for (or even against?) whether the uniform predictor is the right predictor to regularize toward. The important criterion for credit on this problem is to demonstrate understanding and critical thinking.

Programming Assignment

For this assignment, you will run an experiment comparing two different versions of linear classifiers for multiclass classification.

Models

The two models you will implement are perceptron and logistic regression. The multiclass forms of these models are summarized here.

Multiclass Perceptron The multiclass perceptron uses a weight vector for each class, which can conveniently be represented with a matrix $W = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_C\}$. The prediction function is

$$f_{\text{perc}}(\boldsymbol{x}) := \underset{c \in \{1,\ldots,C\}}{\arg\max}\ \boldsymbol{w}_c^\top \boldsymbol{x} = \underset{c \in \{1,\ldots,C\}}{\arg\max}\ \left[W^\top \boldsymbol{x}\right]_c . \tag{8}$$

The multiclass perceptron update rule when learning from example $\boldsymbol{x}_t$, ground-truth label y_t is.

$$\boldsymbol{w}_{y_t} \leftarrow \boldsymbol{w}_{y_t} + \boldsymbol{x}_t \tag{9}$$

$$\boldsymbol{w}_{(f_{\text{perc}}(\boldsymbol{x}))} \leftarrow \boldsymbol{w}_{(f_{\text{perc}}(\boldsymbol{x}))} - \boldsymbol{x}_t \tag{10}$$

Multiclass Logistic Regression Multiclass logistic regression also uses a weight vector for each class, and in fact has the same prediction formula as perceptron.

$$f_{\text{lr}}(\boldsymbol{x}) := \underset{c \in \{1,\ldots,C\}}{\arg\max}\ \boldsymbol{w}_c^\top \boldsymbol{x} = \underset{c \in \{1,\ldots,C\}}{\arg\max}\ \left[W^\top \boldsymbol{x}\right]_c . \tag{11}$$

The key difference from perceptron is that it is built around a probabilistic interpretation:

$$p_{\text{lr}}(y|\boldsymbol{x}; W) := \frac{\exp(\boldsymbol{w}_y^\top \boldsymbol{x})}{\sum_{c=1}^{C} \exp(\boldsymbol{w}_c^\top \boldsymbol{x})} . \tag{12}$$

For data set $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, the regularized negative log likelihood is

$$L(D) = \frac{\lambda}{2}||W||_{\text{F}}^2 + \sum_{i=1}^{n} \log\left(\sum_c \exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)\right) - \sum_{i=1}^{n} \boldsymbol{w}_{y_i}^\top \boldsymbol{x}_i \tag{13}$$

where $||W||_{\mathrm{F}}^2$ is the squared Frobenius norm $\sum_{ij} \boldsymbol{w}_i[j]^2$, and the gradient of the log likelihood is

$$\nabla_{\boldsymbol{w}_c} L = \lambda \boldsymbol{w}_c + \sum_{i=1}^{n} \boldsymbol{x}_i \left(\frac{\exp(\boldsymbol{w}_c^\top \boldsymbol{x}_i)}{\sum_{c'} \exp(\boldsymbol{w}_{c'}^\top \boldsymbol{x}_i)} - I(y_i = c) \right) \tag{14}$$

$$= \lambda \boldsymbol{w}_c + \sum_{i=1}^{n} \boldsymbol{x}_i \left(p_{\mathrm{lr}}(y|\boldsymbol{x}_i; W) - I(y_i = c) \right) \tag{15}$$