

# Machine Learning Fall 2017 Homework 2

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

## General Instructions

You are allowed to work with at most one other student on the homework. With your partner, you will submit only one copy, and you will share the grade that your submission receives. You should set up your partnership on Canvas as a two-person group.

Submit your homework electronically on Canvas. We recommend using LaTeX, especially for the written problems. But you are welcome to use anything as long as it is neat and readable.

Include a README file that describes all other included files. Indicate which files you modified. You are welcome to create additional functions, files, or scripts, and you are also welcome to modify the included interfaces for existing functions if you prefer a different organization.

Since we may work on some of the homework in class, clearly indicate which parts of your submitted homework are work done in class and which are your own work.

Relatedly, cite all outside sources of information and ideas.

## Written Problems

1. Multiclass logistic regression has the form

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

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

- (a) (5 points) You have a batch of  $n$  data points and labels  $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ . 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 (7).
- (b) (5 points) Derive the gradient of the logistic regression likelihood with respect to any one of the  $\mathbf{w}_c$  vectors. It should look very similar to Equation (8) or Equation (9) below (either form is acceptable for full credit). Show and explain the steps you take to derive the gradient.
- (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.

## 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  $W = \{\mathbf{w}_1, \dots, \mathbf{w}_C\}$ . The prediction function is

$$f_{\text{perc}}(\mathbf{x}) := \arg \max_{c \in \{1, \dots, C\}} \mathbf{w}_c^\top \mathbf{x} = \arg \max_{c \in \{1, \dots, C\}} [W^\top \mathbf{x}]_c. \quad (2)$$

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

$$\mathbf{w}_{y_t} \leftarrow \mathbf{w}_{y_t} + \mathbf{x}_t \quad (3)$$

$$\mathbf{w}_{(f_{\text{perc}}(\mathbf{x}))} \leftarrow \mathbf{w}_{(f_{\text{perc}}(\mathbf{x}))} - \mathbf{x}_t \quad (4)$$

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

$$f_{\text{lr}}(\mathbf{x}) := \arg \max_{c \in \{1, \dots, C\}} \mathbf{w}_c^\top \mathbf{x} = \arg \max_{c \in \{1, \dots, C\}} [W^\top \mathbf{x}]_c. \quad (5)$$

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

$$p_{\text{lr}}(y|\mathbf{x}; W) := \frac{\exp(\mathbf{w}_y^\top \mathbf{x})}{\sum_{c=1}^C \exp(\mathbf{w}_c^\top \mathbf{x})}. \quad (6)$$

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

$$L(D) = \frac{\lambda}{2} \|W\|_F^2 + \sum_{i=1}^n \log \left( \sum_c \exp(\mathbf{w}_c^\top \mathbf{x}_i) \right) - \sum_{i=1}^n \mathbf{w}_{y_i}^\top \mathbf{x}_i \quad (7)$$

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

$$\nabla_{\mathbf{w}_c} L = \lambda \mathbf{w}_c + \sum_{i=1}^n \mathbf{x}_i \left( \frac{\exp(\mathbf{w}_c^\top \mathbf{x}_i)}{\sum_{c'} \exp(\mathbf{w}_{c'}^\top \mathbf{x}_i)} - I(y_i = c) \right) \quad (8)$$

$$= \lambda \mathbf{w}_c + \sum_{i=1}^n \mathbf{x}_i (p_{\text{lr}}(y|\mathbf{x}_i; W) - I(y_i = c)) \quad (9)$$

## Experiments

You'll build learning and prediction functions for perceptron and linear regression. You should be able to run an experiment on synthetic data, which will be possible to visualize on the screen, and on real medical data. These experiments are already set up for you in the two iPython notebooks: `synthetic_experiments.ipynb` and `cardio_experiments.ipynb`. The cardiotocography data includes measurements from fetal cardiotocograms that medical experts diagnosed into ten possible diagnosis classes.<sup>1</sup> Both experiments will measure how the perceptron model behavior changes as it sees more examples and how logistic regression behaves as you vary its regularization parameter.

## Tasks

1. (4 points) You should only need to modify `linearclassifier.py`, which has three unfinished functions. The first function you should finish is `linear_predict`, which takes a set of data and a model object as input and outputs the predicted labels for each example. This linear prediction should be done with matrix operations and should take approximately two to five lines of code. This function will be used as the predictor function for both learning algorithms.

The model is a dict that contains only one entry. The key is 'weights', and the value is the weight matrix for the multi-class linear classifier.

<sup>1</sup><https://archive.ics.uci.edu/ml/datasets/Cardiotocography>

2. (4 points) Implement `perceptron_update`, which should run one perceptron learning step. It receives as input **one** example and its label and modifies the `model` object in place. In other words, the function should change the weights of the `model` dictionary to the new weights after applying the perceptron equations. Finally, the function should return whether the perceptron was correct on the given example.

Once you complete this task, you should be able to run the perceptron experiments. The notebooks should generate plots of the training and testing accuracy as you run multiple epochs of online learning.

3. (7 points) Implement `log_reg_nll` (short for *logistic regression negative log-likelihood*), which is a function defined inside `log_reg_train` (short for *logistic regression train*). The function `log_reg_nll` returns a length-two tuple containing (1) the value of the negative log-likelihood—i.e., Equation (7)—and (2) the gradient of the negative log-likelihood with respect to the model weights—i.e., Equation (9).

The first cell of the logistic regression portion in each notebook tests whether your objective function and gradient agree with each other by comparing your provided gradient with a numerically estimated gradient of your objective function. Use that to validate that you implemented the function correctly.

Make sure to read and understand the other code in `log_reg_train`, which uses the function and gradient you compute and passes it into a general purpose numerical optimization tool to perform gradient-based minimization of the learning objective. You technically won't be graded on your understanding of this part for this homework, but it's an important concept in machine learning that will come up again.

Once you complete this task, you should be able to run the logistic regression experiment. The notebooks should plot the training and testing accuracy as the regularization parameter  $\lambda$  changes.