Assignment Three
ECE 4950

- Provide credit to **any sources** other than the course staff that helped you solve the problems. This includes **all students** you talked to regarding the problems.
- You can look up definitions/basics online (e.g., wikipedia, stack-exchange, etc)
- **The due date is 3/22/2017, 23.59.59 Eastern time**.
- Submission rules are the same as previous assignments.

Problem 1 (10 points). Recall that for a feature vector $\bar{X} = (\bar{X}_1, \ldots, \bar{X}_k)$, a linear classifier with weights $\bar{w}^* = (w_0, w_1, \ldots, w_k)$ has decision rule: $\sum_{j=1}^{k} \bar{X}_j w_j \lesseqgtr w_0$. We can append a dummy feature $\bar{X}_0 = -1$, and obtain the new feature vector $\bar{X}^* = (\bar{X}_0, \bar{X}_1, \ldots, \bar{X}_k)$. The decision rule then becomes $\bar{X}^* \cdot \bar{w}^* \lesseqgtr 0$.

Recall the perceptron algorithm from the lectures (scanned notes p.46). Suppose $k = 2$, $n = 4$, and there are four training examples, given by:

training # i	feature $\bar{X}(i)$	label $f(\bar{X}(i))$
1	$(-0.6, 1)$	-1
2	$(-3, -4)$	-1
3	$(3, -2)$	$+1$
4	$(0.5, 1)$	$+1$

1. Start with the weight vector $(0, 0, 1)$ (the x-axis). Implement the perceptron algorithm **by hand**. Go over the data-points in order. Output a table of the following form, which shows the updates you make. We have filled in some entries in the first two rows. You need to add rows until no mistakes happen on any example.

starting weight	example	label	predicted label	new weight
$(0, 0, 1)$	$(-1, -0.6, 1)$	-1	...	...
...	$(-1, -3, -4)$	-1	...	...
...	...	...	...	...

2. Draw a 2-d grid. On this grid, mark the four examples (like we do on the board). Draw the line you obtain as the final result.

Problem (10 points). Recall the log-likelihood function for logistic regression:

$$\ell(\bar{w}^*) = \sum_{i=1}^{n} \log Pr(f(\bar{X}_i) | X_i, \bar{w}^*),$$

where $Pr(f(\bar{X}_i)|X_i, \bar{w}^*)$ is the same as defined in the class for logistic regression, and $f(\bar{X}) \in \{0, 1\}$. Show that $\ell(\bar{w}^*)$ is concave.

Problem 3 (20 points). Suppose we have a classification problem, where each example $\bar{X}$ has k **positive** features. Suppose we restrict weight vectors to be **positive**, and consider the following discriminative function:

$$Pr(f(\bar{X}) = 0|\bar{X}, w_1, \ldots, w_k) = \frac{2}{1 + \exp(\sum_j w_j \bar{X}_j)} = \frac{2}{1 + \exp(\bar{w} \cdot \bar{X})}.$$

1. Find the gradient of $(\bar{X} \cdot \bar{w})^2$ with respect to $\bar{w}$.
2. Write the log-likelihood function $\ell(\bar{w})$ for this problem.
3. Is $\ell(\bar{w})$ concave in the positive quadrant?
4. Compute the gradient of $\ell(\bar{w})$.
5. Write one step of the gradient ascent step to find the maximum likelihood estimate of $\bar{w}$.

Problem 4. (10 points). Recall in linear regression, the labels which are real numbers. Least squares regression minimizes $\sum_{i=1}^{n}(f(\bar{X}(i)) - \bar{X}(i) \cdot \bar{w} - \bar{w}_0)_2^2$. In class we showed that the least squares regression problem is same as the maximum likelihood estimation under Gaussian noise-model, namely where the output is is modeled as $\bar{X} \cdot \bar{w} + \bar{w}_0 + \varepsilon$ for Gaussian ε.

Instead of Gaussian noise, suppose the ε is a Laplace distribution with density:

$$p(\epsilon) = \frac{1}{2\lambda} \exp\left(-\frac{|\varepsilon|}{\lambda}\right), \lambda > 0.$$

1. Show that p is a probability density function over $\mathbb{R}$.
2. Write the log-likelihood function ℓ under this noise model.
3. Show that the maximum of ℓ is in fact the solution to:

$$\min_{\bar{w}, \bar{w}_0} \sum_{i=1}^{n} |f(\bar{X}(i)) - \bar{X}(i) \cdot \bar{w} - \bar{w}_0|.$$

Problem 5 (20 points). We will study the effect of regularization in classification using logistic regression. We play with the Red Wine Quality Dataset.

Please download the datasets `wine-training1.csv`, `wine-training2.csv` and `wine-tasting.csv`. The second training set is a strict subset of the first dataset.

Implement logistic regression in sklearn, using ℓ_2 regularization with regularizer value C in the set $\{0.0000001 \cdot 2^i; i = 0, 1, \ldots, 29\}$. (The regularization paramter is 'C' in sklearn). Use the solver `sag`, which stands for Stochastic Average Gradient. Train on the file `wine-training1.csv` for each regularizer value and test them on `wine-tasting.csv`. Repeat for the training set `wine-training2.csv`.

1. In the same figure, plot the error on the test-set for the classifiers obtained from the two training sets, as a function of C.

2. In the same figure, plot the error on the training-set for the classifiers obtained from the two training sets, as a function of C.

3. Suppose you train the model using `wine-training2.csv`, and obtain the training errors. Which value of C would you choose for the final model? Please explain.

4. Please consider the test errors and describe behaviors such as under-fitting, and over-fitting as a function of C's.

5. For large values of C (weak regularization), which of the two training data-sets yields a smaller error? Please explain.