Introduction to Machine Learning
Homework 3: Logistic Regression*

Prof. Linda Sellie

For any calculations, you *must* show your work.

1. How does the logistic function change when w_0 changes? You can just run some simulations and describe what you notice. (Or state mathematically what happens)

2. How does the logistic function change if you use $\mathbf{w}' = 2\mathbf{w}$ instead of $\mathbf{w}$? You can just run some simulations and describe what you notice. (Or state mathematically what happens.) Try increasing by larger factors than the number 2. Use your observations to argue why a solution with large weights can cause logistic regression to overfit.

3. Suppose you are in the middle of training a logistic classifier on a data set (below) where the current coefficients are: $\mathbf{w}^T = [0.66, -2.24, -0.18]$.

   In the table below $h_{\mathbf{w}}(x) = \frac{1}{1+e^{-(w_0+w_{1:k}^T\mathbf{x})}}$

	x_1	x_2	$h_{\mathbf{w}}(x)$	y
1	0.49	0.09	0.389	0
2	1.69	0.04	0.042	0
3	0.04	0.64	0.613	0
4	1.	0.16	0.167	0
5	0.16	0.09	0.572	1
6	0.25	0.	0.526	1
7	0.49	0.	0.393	1
8	0.04	0.01	0.638	1

   (a) What is the equation for the decision boundary?

   (b) For a decision boundary of 0.5, create the confusion matrix.

   (c) Plot the points on a graph and draw the decision boundary (I would suggest using some plotting library and an image editor)

   (d) For the data set above, what is the FPR?

   (e) For the data set above, what is the TPR?

   (f) What is the accuracy?

   (g) What is the recall?

   (h) What is the precision

*Many of these questions are from Prof. Rangan.

(i) Write the likelihood function $L(\mathbf{w})$ for the training examples above.

(j) In logistic regression, we are trying to maximize the log-likelihood

$$\ell(\mathbf{w}) = \sum_{i=1}^{N} (y^{(i)} \ln(h(\mathbf{x}^{(i)})) + (1 - y^{(i)}) \ln(1 - h(\mathbf{x}^{(i)})))$$

which is the same as minimizing the error function

$$- \left(\sum_{i=1}^{N} (y^{(i)} \ln(h(\mathbf{x}^{(i)})) + (1 - y^{(i)}) \ln(1 - h(\mathbf{x}^{(i)}))) \right)$$

This quantity is sometimes called the *cross-entropy* of the classifier on the dataset. Using the initial weights, what is the cross-entropy of the classifier on the given training set?

(k) Given $\mathbf{w}$ as described above and $\mathbf{w}' = (1.33, -2.96, -2.77)^T$, which is more likely to be the correct decision boundary given access only to the data above.

(l) Perform one step of gradient ascent using the $\mathbf{w}$ above and learning rate 0.1

(m) How did the data points near the decision boundary contribute to the new value of $\mathbf{w}$?

(n) How did the data points which were correctly classified and far away from the decision boundary contribute to the new value of $\mathbf{w}$?

(o) How did incorrectly classified points contribute to the new value of $\mathbf{w}$?

(p) Compute the cross-entropy (error) with the weights you computed in step **??**. Did the cross-entropy (error) go up or down after one iteration of the gradient ascent? Is this what you expected? Why or why not?

4. Suggest possible label variables (target/response variables) and features (predictors) for the following classification problems. For each problem, indicate how many classes there are. There is no single correct answer.

   (a) Given an audio sample to detect the gender of the voice.

   (b) A electronic writing pad records the motion of a stylus, and it is desired to determine which letter or number was written. Assume a segmentation algorithm is already run, which indicates very reliably the beginning and end time of the writing of each character.

5. Regularization:

   (a) Add lasso regularization to the log-likelihood function for logistic regression

   (b) Add ridge regularization to the log-likelihood function for logistic regression

   (c) Determine the derivative of the log-likelihood function for logistic regression with ridge regularization.

   (d) Implement logistic regression with ridge regularization. Submit a screenshot of the code and text boxes as part of your written solution.

   - Add the ridge regularization to your programming assignment for logistic regression

- Using 5-fold cross-validation, find the optimal λ. Did the regularization help? How did the regularization affect the error in the training data? How did it affect the error on the validation set?

6. A data scientist is hired by a political candidate to predict who will donate money. The data scientist decides to use two predictors for each possible donor:

   - x_1 = the income of the person(in thousands of dollars), and
   - x_2 = the number of websites with similar political views as the candidate, the person, follows on Facebook.

   To train the model, the scientist tries to solicit donations from a randomly selected subset of people and records who donates or not. She obtains the following data:

Income (thousands \$), $x_1^{(i)}$	30	50	70	80	100
Num websites, $x_2^{(i)}$	0	1	1	2	1
Donate (1=yes or 0=no), $y^{(i)}$	0	1	0	1	1

   (a) Draw a scatter plot of the data labeling the two classes with different markers.

   (b) Find a linear classifier that makes at most one error on the training data. The classifier should be of the form,

$$\hat{y}_i = \begin{cases} 1 & \text{if } z^{(i)} > 0 \\ 0 & \text{if } z^{(i)} < 0, \end{cases} \qquad z^{(i)} = \mathbf{w}^{\mathsf{T}}\mathbf{x}^{(i)}$$

   What is your classifier's weight vector $\mathbf{w}$?

   (c) Now consider a logistic model of the form,

$$P(y^{(i)} = 1|\mathbf{x}^{(i)}) = \frac{1}{1+e^{-z^{(i)}}}, \quad z^{(i)} = \mathbf{w}^{\mathsf{T}}\mathbf{x}^{(i)}$$

   Using $\mathbf{w}$ from the previous part, which sample i is the *least* likely (i.e. $P(y^{(i)}|\mathbf{x}^{(i)})$ is the smallest). If you do the calculations correctly, you should not need a calculator.

   (d) Now consider a new set of parameters

$$\mathbf{w}' = \alpha\mathbf{w},$$

   Where $\alpha > 0$ is a positive scalar. Would the new parameters change the values $\hat{y}$ in part (b)? Would they change the likelihoods $P(y_i|\mathbf{x}_i)$ in part (c)? If they do not change, state why. If they change, qualitatively describe the change as a function of α.

7. (Do not turn in) [1]

   A multivariate logistic regression model has been built to predict the propensity of shoppers to perform a repeat purchase of a gift they are given. The descriptive features used by the model are the age of the customer, the socioeconomic band to which the customer belongs (a, b, or c), the average amount of money the customer spends on each visit to the shop, and the average number of visits the customer makes to the shop per week. The marketing department uses this model to determine who should receive the free gift.

[1] This is a modified question from Machine Learning For Predictive Data Analytics

(a) The weights in the trained model are shown in the following table:

Feature	Weight
Intercept (w_0)	-3.82398
AGE	-0.02990
SOCIOECONOMIC BAND B	-0.09089
SOCIOECONOMIC BAND C	-0.19558
SHOP VALUE	0.02999
SHOP FREQUENCY	0.74572

Create the coefficient vector $\mathbf{w}$.

(b) Rewrite the following data matrix using the dummy encoding to work with the coefficients in the previous question.

ID	AG	SOCIOECONOMIC BAND	SHOP FREQUENCY	SHOP VALUE
1	56	b	1.60	109.32
2	21	c	4.92	11.28
3	48	b	1.21	161.19
4	37	c	0.72	170.65
5	32	a	1.08	165.39

(c) Use the model to make predictions for the data matrix you created in the previous question.

(d) It is recommended that all continuous descriptive features be scaled in building logistic regression models. In this question, the continuous features were normalized to the range $[-1, 1]$ using min/max normalization (also called range normalization). The following table shows a data quality report for the dataset used to train the model described above.

Feature	N	% Missing	min value	mean	max value	std. dev
Age	5,2000	6	18	32.7	63	12.2
SHOP FREQUENCY	5,2000	0	0.2	2.2	5.4	1.6
SHOP VALUE	5,2000	0	5	101.9	230.7	72.1

Feature	N	% Missing	# categories	mode
SOCIOECONOMIC BAND	5,2000	8	3	a
REPEAT PURCHASE	5,2000	0	2	no

On the basis of the information in this report, all continuous features were normalized using min/max normalization (aka range normalization).

After applying these data preparation operations, a logistic regression model was trained to give the weights shown in the following table.

Feature	Weight
Intercept (w_0)	0.6679
AGE	-0.5795
SOCIOECONOMIC BAND B	-0.1981
SOCIOECONOMIC BAND C	-0.2318
SHOP VALUE	3.4091
SHOP FREQUENCY	2.0499

For this question, if we have any missing values, we will replace them using mean imputation for continuous features and mode imputation for categorical features.

Use this model to make predictions for each query instance shown in the following table (question marks refer to missing values).

ID	AG	SOCIOECONOMIC BAND	SHOP FREQUENCY	SHOP VALUE
1	38	a	1.90	165.39
2	56	b	1.60	109.32
3	18	c	6.00	10.09
4	?	b	1.33	204.62
5	62	?	0.85	110.50