Machine Learning: Assignment 4

Due on Wed, Feb 29, 2023 at 11:00am

Instructor: Dr. Anh Nguyen

Note: Late assignments will NOT be graded at all unless you have been granted extensions in advance (see the policy in syllabus).
There are 6 problems for a total of 100 points.

Submission:
- Please write the answers to Problem 2, 3, 4, 5, 6 in a typeset format (i.e., all your answers must be typed up using Word or Latex) and then generate a single PDF named **A4_problem2-6.pdf**. Handwritten work will not be graded.
- For Problem 1, generate **A4_problem1.ipynb**.
- Please zip up both 2 files (A4_problem1.ipynb, A4_problem2-6.pdf) into a single zip file and upload that to Canvas.

Problem 1

50 points Complete the 3 incomplete functions in this Google Colab to train a Logistic Regression binary classifier on the UCI breast cancer dataset.
Specifically, you are asked to complete the following three functions:

- _sigmoid_fn
- get_loss
- get_gradients

Upon completion, your classifier is expected to obtain ≥ 94% accuracy.
Submission Download your edited ipynb and rename it as **A4_problem1.ipynb**

Problem 2

10 points Explain in 3-5 sentences why logistic regression is a *discriminative* classifier as opposed to Naive Bayes, which is a *generative* classifier. In other words, what makes logistic regression *discriminative* and what makes Naive Bayes *generative*?
Notes: Write down mathematics equations, if necessary, to support your answer.

Logistic Regression is a discriminative classifier because it assumes a functional form P(y|x) to estimate parameters P(y|x) directly from the training data and requires no independence assumptions. Naïve Bayes is a generative classifier because it is based on the functional form P(x|y),P(x) directly from the training data and using Bayes rule to calculate P(y|x) this requires an assumption of conditional independence.

Problem 3

10 points The prediction rule for a logistic-regression, binary classifier is if $P(y = 1|\boldsymbol{x}) > P(y = 0|\boldsymbol{x})$ then output 1 otherwise, output 0.
Assume that our $\boldsymbol{w} \in R^2$ (i.e., there are two weights w_1 and w_2 only) and there is a bias $b \in R$. The label $y \in \{0, 1\}$. The input features $\boldsymbol{x} \in R^2$.
Question: Derive the decision boundary equation for this classifier.
Hints: The decision boundary of this logistic regression classifier is a line at the point when

$$P(y = 1|\boldsymbol{x}) = P(y = 0|\boldsymbol{x})$$

From the above equation, continue deriving step-by-step (with justifications) to arrive at the equation for the linear decision boundary (of this logistic regression classifier) separating two classes ($y = 0$ and $y = 1$).

First find the sigmoid form of the conditional probability

$$P(y=1|x) = \frac{P(Y=1)P(X|Y=1)}{P(Y=1)P(X|Y=1) + P(Y=0)P(X|Y=0)}$$

$$\frac{1}{1 + \frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}}$$

$$\frac{1}{1 + e^{\ln\left(\frac{P(Y=0)P(X|Y=0)}{P(Y=1)P(X|Y=1)}\right)}}$$

$$\frac{1}{1 + e^{\ln\left(\frac{1-\pi}{\pi}\right) + \sum_i^N \ln\left(\frac{P(X_i|Y=0)}{P(X_i|Y=1)}\right)}}$$

$$\frac{1}{1 + e^{\ln\left(\frac{1-\pi}{\pi}\right) + \sum_i^N \left(\frac{\mu_{i0} - \mu_{i1}}{\sigma_i^2} \mathrm{X_i} + \frac{\mu_{i0}^2 - \mu_{i1}^2}{2\sigma_i^2}\right)}}$$

$$where\ b = w_0$$

$$P(y=1|x) = \frac{1}{1 + e^{(w_0 + \sum w_i^t x_i)}}$$

$$P(y=0|x) = \frac{e^{w_0 + \sum w_i^t x_i}}{1 + e^{(w_0 + \sum w_i^t x_i)}}$$

The decision boundary is found where

$$P(y=1|x) = P(y=0|x)$$

Plugging in $P(y=1|x)$, *and* $P(y=0|x)$

$$\frac{P(Y=0|X)}{P(Y=1|X)} = \frac{\frac{e^{\sum w^t x + b}}{1 + e^{(\sum w^t x + b)}}}{\frac{1}{1 + e^{(\sum w^t x + b)}}}$$

$$\frac{P(Y=0|X)}{P(Y=1|X)} = e^{b + \sum w^t x}$$

we take the ln of both sides to eliminate the e

$$ln\frac{P(Y=0|X)}{P(Y=1|X)} = b + \sum w^t x$$

$$wx_1 + wx_2 + b = 0$$

Problem 4

10 points Given the classifier in Problem 2, derive the full conditional log likelihood function $L(\boldsymbol{w})$ that we want to maximize (i.e. MCLE not MAP). We have N examples (i.e., pairs of $(\boldsymbol{x}^i, y^i)$) in the training set.

$$L(w) = \log_e \prod P(y^l | x^l, w)$$

$$\sum_l^N \ln P(y^l | x^l, w)$$

$$l(w) = \sum Y^l \ln\left(P(y=1|x^l,w)\right) + (1-Y^l)\ln\left(P(y=0|x^l,w)\right)$$

$$\sum Y^l \ln\left(\frac{P(y=1|x,w)}{P(y=0|x,w)}\right) + \ln\left(P(y^l=0|x,w)\right)$$

$$\sum Y^l \left(w_0 + \sum w_i x_i^l\right) - \ln\left(1 + \exp\left(w_0 + \sum_i^n w_i x_i^l\right)\right)$$

Problem 5

10 points Extend the given classifier in Problem 2 to a single-label, 3-way classifier, i.e., given $\boldsymbol{x} \in \mathbb{R}^2$, predict one label for $y \in \{0, 1, 2\}$.
Hint: How many set of weights and biases are there in this case?

Question: Write down the full definition of the three below functions (i.e., in terms of exp and the parameters w_1, w_2,...so that someone can plug in the parameters and compute the output probabilities).

When Y is not Boolean $y \in \{y_1 \dots y_R\}$

$$k < R;\ P(Y = y_k | x) = \frac{e^{w_{k0} + \sum_{j=i}^{R-1} w_{ki} X_i}}{1 + \sum_{j=i}^{R-1} e^{w_{j0} + \sum_{i=i}^{n} w_{ij} X_i}}$$

$$k = R;\ P(Y = y_R | X) = \frac{1}{1 + \sum_{j=i}^{R-1} e^{w_{j0} + \sum_{i=i}^{n} w_{ij} X_i}}$$

$P(y = 0 | \boldsymbol{x}) =$

$$\frac{\exp(w_{00} + \sum w_{0i} x_i)}{1 + \sum_{j=i}^{R-1} exp(w_{j0} + \sum w_{ji} x_i)} =$$

$$\frac{e^{b_1 + w_{11} x_1 + w_{12} x_2}}{1 + e^{b_1 + w_{11} x_1 + w_{12} x_2} + e^{b_2 + w_{21} x_1 + w_{22} x_2}}$$

$P(y = 1 | \boldsymbol{x}) =$

$$\frac{\exp\left(w_{10}+\sum w_{1i}x_i\right)}{1+\sum_{j=i}^{R-1} exp(w_{j0}+\sum w_{ji}x_i)} =$$

$$\frac{e^{b_1+w_{11}x_1+w_{12}x_2}}{1+e^{b_1+w_{11}x_1+w_{12}x_2}+e^{b_2+w_{21}x_1+w_{22}x_2}}$$

$P\left(y=2|\boldsymbol{x}\right) =$

$$\frac{1}{1+\sum_{j=i}^{R-1} exp(w_{j0}+\sum w_{ji}x_i)} =$$

$$\frac{1}{1+e^{b_1+w_{11}x_1+w_{12}x_2}+e^{b_2+w_{21}x_1+w_{22}x_2}}$$

Problem 6

10 points Given the classifier in Problem 2, extend the classifier to a single-label, 3-way **softmax** classifier. That is, given $\boldsymbol{x} \in \mathrm{R}^2$, predict one label for $y \in \{0, 1, 2\}$. Yet, use the **softmax** function instead of the **sigmoid**.

Question: Write down the full definition of the three below functions (i.e., in terms of exp and the parameters w_1, w_2,...so that someone can plug in the parameters and compute the output probabilities).

The SoftMax function is shown, *where z is the input vector*, $b_i + w_i x_i$, as

$$P(y=c|x,W,b) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

We have 3 probabilities, so three denominator vectors,

$P\left(y=0|\boldsymbol{x}\right) =$

$$\frac{e^{b_1+w_{11}x_1+w_{12}x_2}}{e^{b_1+w_{11}x_1+w_{12}x_2}+e^{b_2+w_{21}x_1+w_{22}x_2}+e^{b_3+w_{31}x_1+w_{32}x_2}}$$

$P\left(y=1|\boldsymbol{x}\right) =$

$$\frac{e^{b_2+w_{21}x_1+w_{22}x_2}}{e^{b_1+w_{11}x_1+w_{12}x_2}+e^{b_2+w_{21}x_1+w_{22}x_2}+e^{b_3+w_{31}x_1+w_{32}x_2}}$$

$P\left(y=2|\boldsymbol{x}\right) =$

$$\frac{e^{b_3+w_{31}x_1+w_{32}x_2}}{e^{b_1+w_{11}x_1+w_{12}x_2}+e^{b_2+w_{21}x_1+w_{22}x_2}+e^{b_3+w_{31}x_1+w_{32}x_2}}$$