# STAT 543/641 – Winter 2019 – Homework #2

Due Wednesday, March 20, 2019

**Notation:** Suppose $(x_1, y_1, \ldots, (x_n, y_n) \in \mathbb{R} \times \{0, 1\}$. Set

$$\begin{aligned}
\sigma(t) &= \frac{1}{1+e^{-t}} \\
p_i(a,b) &= \sigma(a+bx_i)^{y_i}\big(1-\sigma(a+bx_i)\big)^{1-y_i} \\
\ell_i(a,b) &= -\log p_i(a,b) \\
&= -y_i \log \sigma(a+bx_i) - (1-y_i)\log\big(1-\sigma(a+bx_i)\big) \\
\ell(a,b) &= \sum_{i=1}^{n} \ell_i(a,b)
\end{aligned}$$

1. In this problem, we establish a sufficient condition for the uniqueness of maximum likelihood estimates for univariate logistic regression coefficients, assuming such estimates exist.

   (a) Prove:
   $$\sigma'(x) = \sigma(x)\big(1-\sigma(x)\big) \tag{$*$}$$
   Conclude that $\sigma'(x) > 0$ for all $x$.

   (b) Prove that
   $$\nabla \ell_i(a,b) = (\sigma(a+bx_i) - y_i)\begin{bmatrix} 1 \\ x_i \end{bmatrix} \qquad \text{(Hint: Use } (*)\text{.)}$$
   and that
   $$\nabla^2 \ell_i(a,b) = \sigma'(a+bx_i)\begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix}.$$
   Deduce that $\nabla^2 \ell_i(a,b)$ is positive-semidefinite but not positive-definite, making $\ell_i(a,b)$ convex but not strictly convex.

(c) Find a basis of the nullspace $N(\nabla^2 \ell_i(a,b))$ of $\nabla^2 \ell_i(a,b)$ whose elements do not depend on $a$ and $b$.

(d) Suppose that there are indices $i$ and $j$ such that $x_i \neq x_j$. Prove that

$$\bigcap_{i=1}^{n} N(\nabla^2 \ell_i(a,b)) = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}.$$

(Hint: Use (c).)

(e) Suppose that there are indices $i$ and $j$ such that $x_i \neq x_j$. Show that $\nabla^2 \ell(a,b)$ is positive-definite and, hence, that $\ell(a,b)$ is strictly convex.

(f) Conclude that if that there are indices $i$ and $j$ such that $x_i \neq x_j$, then maximum likelihood estimates for $\widehat{a}$ and $\widehat{b}$ are unique if they exist.

2. In this problem, we establish a sufficient condition for the existence of maximum likelihood estimates for univariate logistic regression coefficients.

   Consider fitting a univariate logistic regression model to a dataset $(x_1, y_1), \ldots, (x_n, y_n)$ satisfying

$$x_1 < x_2 < \cdots < x_n.$$

(a) Prove that $\ell_i(a,b) > 0$ for all $(a,b) \in \mathbb{R}^2$.

(b) Let

$$H_i = \left\{ \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \in \mathbb{R}^2 : \lim_{t \to \infty} \ell_i(tv_1, tv_2) = \infty \right\}$$

Find a vector $\boldsymbol{w} \in \mathbb{R}^2$ such that $H_i = H(\boldsymbol{w}_i)$, where

$$H(\boldsymbol{w}_i) = \left\{ \boldsymbol{v} \in \mathbb{R}^2 : \boldsymbol{v} \cdot \boldsymbol{w}_i > 0 \right\}.$$

(c) Let $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \mathbb{R}^2 - \{\boldsymbol{0}\}$. Show that

$$H(\boldsymbol{u}) \cup H(\boldsymbol{v}) \cup H(\boldsymbol{w}) = \mathbb{R}^2 - \{\boldsymbol{0}\}$$

if and only there are $a, b > 0$ such that $-\boldsymbol{u} = a\boldsymbol{v} + b\boldsymbol{w}$.

(d) Consider the following condition on a triple of indices $(i, j, k)$:

$$y_i = y_k = 0 \text{ and } y_j = 1 \qquad \textit{or} \qquad y_i = y_k = 1 \text{ and } y_j = 0 \tag{$\dagger$}$$

Suppose $1 \leq i < j < k \leq n$. Prove that $(i, j, k)$ satisfies $(\dagger)$ if and only if

$$H_i \cup H_j \cup H_k = \mathbb{R}^2 - \{\boldsymbol{0}\}.$$

(e) Suppose that $(i, j, k)$ is an increasing sequence of indices that satisfies (†). Prove that, for all $K > 0$, the set

$$S_K := \{(a, b) : \ell(a, b) \leq K\}$$

contains no ray from the origin, i.e., no set of the form $\{t\boldsymbol{v} : t \geq 0\}$ where $\boldsymbol{v} \in \mathbb{R}^2 - \{\mathbf{0}\}$.

(f) Use the following facts to deduce that $S_K$ is bounded for all $K > 0$.

i. If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, then $\{\boldsymbol{x} \in \mathbb{R}^n : f(\boldsymbol{x})| \leq K\}$ is a convex set.

ii. If $C$ is a convex set that contains no ray, then $C$ is bounded.

(g) Let $K > 0$ be such that $S_K$ is nonempty and let

$$m = \inf_{(a,b) \in S_K} \ell(a, b).$$

Explain why there exists a point $(\widehat{a}, \widehat{b}) \in S_K$ such that $\ell(\widehat{a}, \widehat{b}) = m$ and why $m$ is, in fact, the global minimum of $\ell$.

(h) Prove that $(\widehat{a}, \widehat{b})$ is the unique point at which $\ell$ takes on its minimum value.

3. Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be random samples from normally distributed populations with means $\mu_X$ and $\mu_Y$ and variances $\sigma_X^2$ and $\sigma_Y^2$, respectively. Let $S_X^2$ and $S_Y^2$ be the standard unbiased estimators of $\sigma_X^2$ and $\sigma_Y^2$, respectively. Show that

(a) Suppose $\sigma_X^2 = \sigma_Y^2$ and write $\sigma^2$ for this common value. Show that

$$S^2 := \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

is an unbiased estimator of $\sigma^2$. It's called the *pooled variance estimator*.

(b) Suppose, in addition to having common variance, that the $X_i$ are independent of the $Y_i$. What is the distribution of

$$\frac{(m+n-2)S^2}{\sigma^2}?$$

What is the variance of $S^2$?

(c) (Do not hand in.) Generalize these results from the case of $K = 2$ populations to that of an arbitrary $K$. Compare with equation (4.15) in [1].

(d) (Do not hand in.) Can you prove analogous results with covariance matrices in place of scalar variances?

4. Use whatever software package you want to do this problem. I recommend `R` or `python` with `sklearn`, though.

Load the file `dataset_1.csv`, containing synthetic data for a binary classification problem with two numerical features. These features are in the first two columns of the file; the target is in the third column.

(a) Make a scatter plot of the data, using different colors or marker shapes to distinguish class labels.

(b) Fit the data to a logistic regression model and to a Gaussian naïve Bayes model. Plot the decision boundaries. Do you expect these classifiers to yield satisfactory results? (If your decision boundary looks strange, note the relative sizes of the two classes.)

(c) Write $X_1$ and $X_2$ for the two features. Augment the dataset with three new features:

$$X_3 := X_1^2, \quad X_4 := X_1 X_2, \quad X_5 := X_2^2$$

(d) Fit the data to a logistic regression model and to a Gaussian naïve Bayes model.

(e) Let

$$c_0 + c_1 X_1 + c_2 X_2 + c_3 X_3 + c_4 X_4 + c_5 X_5 = 0.$$

be the decision boundary. Since it's graph is a hyperplane in $\mathbb{R}^5$, we can't plot it. We can, however, plot its inverse image in $\mathbb{R}^2$ under the map

$$(x_1, x_2) \mapsto (x_1, x_2, x_1^2, x_1 x_2, x_2^2),$$

i.e., the plane curve

$$c_0 + c_1 x_1 + c_2 x_2 + c_3 x_1^2 + c_4 x_1 x_2 + c_5 x_2^2 = 0.$$

Plot the decision boundary. Comment on your results. (If it's too much trouble to get the equation of the decision boundary from whatever implementation of Gaussian naïve Bayes you're using, don't worry about it.)

(f) Repeat with the dataset in `dataset_2.csv`, augmenting your dataset with all the monomials in $X_1$ and $X_2$ of degrees 2 and 3, for a total of nine features.

5. Load the file `dataset_3.csv`, containing synthetic data for a four class classification problem with two numerical features. These features are in the first two columns of the file; the target is in the third column. We will use a *one-versus-rest* approach to construct a four class classifier out of four binary classifiers.

   (a) Make a scatter plot of the data, using different colors or marker shapes to distinguish class labels.

   (b) Split the dataset into training and testing subsets. Let $\boldsymbol{y}_{tr}$ and $\boldsymbol{y}_{te}$ be their target vectors.

   (c) Let $k_0 \in \{0, 1, 2, 3\}$. Construct a new training and testing target vectors, $\boldsymbol{y}_{tr,k}$ and $\boldsymbol{y}_{te,k}$, vector by replacing the three class labels $k \neq k_0$ with 4. Fit the resulting training set to a logistic regression model.

   (d) Test your four models on their corresponding testing sets: For each test observation, $x$, your models will give estimates for $p(k|x)$. Assign each test observation to the class label, $k$, for which this probability is the highest. What is the overall relative accuracy of your classifier on the test set?

   (e) (Bonus) Can you plot the decision boundaries?

# References

[1] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *Introduction to Statistical Learning Theory with Applications in R*, `http://www-bcf.usc.edu/~gareth/ISL/`.