ISYE6740 Midterm
September 13, 2018
Solutions; Comments

This exam was not meant to see how many little facts or formulas you had memorized.
It did not ask you just to repeat facts or answers that you had given before.
It was meant to see whether you could use what you had learned, maybe in slightly different situations.

Models of Relationships between Variables.

A very general form of a model of a statistical relationship between two variables (or sets of variables) is

$$h(y) = f(x; \theta, \epsilon),$$

where y and x are observable, θ is unobservable (usually assumed to be a constant), and ϵ is an unobservable random variable. In the general regression model, y is a numeric variable, and in the classification problem, y is a categorical variable denoting the class or group to which an observational unit belongs.

1. **Linear Models**

   A common form of this model is the linear regression model, in which we usually use "β" in place of "θ". In the univariate multiple linear regression model, $h(y) = y$ (a scalar), x and β are vectors of the same order, and $f(x; \theta, \epsilon) = x^{\mathrm{T}}\beta + \epsilon$. (Notice, of course, $x^{\mathrm{T}}\beta = \beta^{\mathrm{T}}x$, so we might write it either way.)

   After we have some data, we often write the model in the form

   $$y = X\beta + \epsilon,$$

   where y, β, and ϵ are vectors and X is a matrix. Let n be the number of observations and p be the number of columns in X.

   (a) The first thing we do usually is fit the model.

   i. **How do we fit the model by least squares?**

   To use least squares to fit a model, need data. Then we find values of the parameters to minimize the sum of squares of the residuals for the data.
   For the model written as $y = x^{\mathrm{T}}\beta + \epsilon$, with data $(y_1, x_1), \ldots, (y_n, x_n)$, it means finding $\widehat{\beta}$ so that

   $$\sum_{i=1}^{n} \left(y_i - x^{\mathrm{T}}\widehat{\beta}\right)^2$$

   is minimized.
   For the model written as $y = X\beta + \epsilon$, with data y and X, it means finding $\widehat{\beta}$ so that

   $$(y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta})$$

   is minimized.

   ii. **How do we fit the model by least absolute values?**

   We find values of the parameters to minimize the sum of absolute values of the residuals for the data.
   For the model written as $y = x^{\mathrm{T}}\beta + \epsilon$, with data $(y_1, x_1), \ldots, (y_n, x_n)$, it means finding $\widehat{\beta}$ so that

   $$\sum_{i=1}^{n} |y_i - x^{\mathrm{T}}\widehat{\beta}|$$

   is minimized.

iii. **To fit the model by maximum likelihood, what else do we need to know, and how do we fit it?**

To fit by maximum likelihood, we also need to know the probability distribution of the observables.
We then form the likelihood function of the parameter. Finally, we find the value of the parameter that yields the maximum. That value is the MLE.

(b) Continuing with the linear model, $y = X\beta + \epsilon$, ...
Now, assume that the elements of the error ϵ are independently distributed as $\mathrm{N}(0, \sigma^2)$, **and** that we use least-squares as the criterion of fitting, and that $\widehat{\beta}$ is the least-squares estimator of β.

i. **What is the distribution of $\widehat{\beta}$?**

If the ϵs are independently distributed as $\mathrm{N}(0, \sigma^2)$, then the ys are independently distributed as $\mathrm{N}(X\beta, \sigma^2)$.
Because $\widehat{\beta}$ is a linear combination of the ys, $\widehat{\beta}$ has a normal distribution.
The mean of $\widehat{\beta}$ is

$$\begin{aligned} \mathrm{E}(\widehat{\beta}) &= \mathrm{E}\left((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y\right) \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{E}(y) \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X\beta \\ &= \beta. \end{aligned}$$

The variance of $\widehat{\beta}$ is

$$\begin{aligned} \mathrm{V}(\widehat{\beta}) &= \mathrm{V}\left((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y\right) \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{V}(y)X(X^{\mathrm{T}}X)^{-1} \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}\sigma^2 \\ &= (X^{\mathrm{T}}X)^{-1}\sigma^2. \end{aligned}$$

ii. **What is the least-squares estimator of the variance σ^2 of ϵ? Be explicit.**

This is not as straightforward as it may appear. Actually, any value of σ^2 cancels out of the sum of squared residuals! The least squares criterion is used indirectly.
The "least squares" estimator of σ^2 is formed from the sum of squares of the least square residuals, r_i. We scale the sum so that it is unbiased.
The "answer" here depends on that old annoyance: does the model include the intercept or have the ys and Xs been centered.
In the most common notation, the data have been centered ant p is the number of X variables (that is, there is no column of 1's in the X matrix. Therefore, the divisor that yields an unbiased estimator is $n - p - 1$:

$$\widehat{\sigma^2} = \frac{1}{n-p-1}\sum_{i=1}^{n} r_i^2 = \mathrm{MSE}.$$

(c) If y is an indicator of a class or group, and the columns of X represent features for classification, the linear regression model can be used as a classifier.

i. **Suppose that there are only two groups, that is, y takes the values 1 or 2. Describe how you would use the linear regression model as a classifier.**

A simple way is just to use the response values as if they were regular numeric values, and perform the regression in the usual way.
In the fitted model, a response less than 1.5 is assigned to group 1, and a value greater than 1.5 is assigned to group 2. A fitted response of 1.5 can be assigned either way.

ii. **Give two reasons why the linear regression model is usually not a good classifier, even if there are only two groups.**

In many classification problems, the boundary between the groups is no linear.
The model does not make sense for values of the predictors that are very large or very small. The predicted response just becomes larger or smaller without bound, although the response is either 1 or 2.

2. **Logistic Models**

The logistic model is another special form of the general model for a function of an observable response, y, an observable set of features x, some unobservable parameters θ, and a function of some unobservable random variable, that I wrote on the first page,

$$h(y) = f(x; \theta, \epsilon).$$

(a) **In the general form above, what is the specific form of $h(y)$ in the logistic model?**

In the standard form of the logistic model, the quantity modeled is the probability that the response takes on a particular group indicator, so $h(y)$ is $\Pr(Y = k|x)$.

(b) **For the form of $h(y)$ from the previous question, what is the specific form of $f(x; \theta, \epsilon)$ in the logistic model?** (The random component ϵ is somewhat unusual. Don't worry about it now; it will come up in the next question.)

Because we are modeling a probability, the right-hand side should always be between 0 and 1. There are several ways we could form a meaningful expression in that range.

The logistic regression model uses a generalized linear form,

$$\frac{\mathrm{e}^{\beta^{\mathrm{T}} x}}{1 + \mathrm{e}^{\beta^{\mathrm{T}} x}}.$$

(c) There is a random component in the logistic model, but we don't write it explicitly as ϵ. It is buried in the probability distribution of the response. **What is that distribution?** (Either just give the distribution's name *and* expression(s) for any parameter(s) of the distribution — or describe the distribution clearly.)

The model specifies a *probability*

There is no additive "error" term. Instead, there is a probability distribution for Y. It is a two-point distribution. (It's called a Bernoulli distribution, and the sum of several independently- and identically-distributed Bernoullis is called a binomial distribution.)

3. **KNN Classifiers**

Suppose we are given a training set in a binary classification problem, and now we want to classify a new observation with features x_0. **Describe how you would classify that observation.**

In a KNN classifier, a value of K is chosen (or given).

To classify an observation with predictor value of x_0, the K observations in the training set whose predictors are closest to x_0 are identified. Of those, the response that occurs most frequently is chosen as the predicted response for the observation with predictor value x_0.

In case of a tie, any rule for breaking the tie may be used.

4. **LDA and QDA as Classifiers**

(a) In both LDA and QDA, the underlying model is a mixture of normal distributions with different means. **Exactly how do the models differ in LDA and QDA?**

In the underlying model for LDA, the variances of all of the normal distributions are the same. In QDA, they may be different.

(b) The difference in the models leads to a major difference in the results of LDA and QDA. **Specifically, how do the results differ; that is, in a very *practical* sense, what is the difference?**

The decision boundaries in LDA are linear in x, whereas in QDA, they are quadratic.

(c) Now, let's see if you *understand* LDA and QDA.

We'll develop an analogous procedure under a different underlying model.

Assume a random response $Y = 1$ or 2, and an associated feature X, such that X has a "double exponential" distribution (also called a "Laplace" distribution) with a mean and scale that depend on whether $Y = 1$ or 2. Specifically, for $Y = k$, X has the probability density function

$$f_{X|Y=k}(x; \mu_k, \theta_k) = \frac{1}{2\theta_k} \mathrm{e}^{-|x-\mu_k|/\theta_k}.$$

Notice that $f_{X|Y=k}(x; \mu_k, \theta_k)$ could be written in standard notation equivalently as $f_{X|Y=k}(x|Y = k)$. (Notice also that this notation is slightly different from that used in the text; however, it is consistent with that in the text and what I have used in class.)

In the overall mixture population, the probability that $Y = k$ is π_k (the "prior probability").

i. **Use Bayes theorem to write an expression for** $\Pr(Y = k|x)$**.**

First, some background on the procedure:
For two random variables Y and X, there are 5 probability density functions (or probability functions):
2 marginals: $f_Y(y)$ and $f_X(x)$
2 conditionals: $f_{Y|x}(y)$ and $f_{X|y}(x)$
1 joint: $f_{Y,X}(y, x)$
They are related in these ways:
$f_{Y,X}(y, x) = f_X(x)f_{Y|x}(y) = f_Y(y)f_{X|y}(x)$ (This is the basis for "Bayes theorem".)
$f_Y(y) = \int f_{Y,X}(y, x)\mathrm{d}x$ and $f_X(x) = \int f_{Y,X}(y, x)\mathrm{d}y$
Now, what does the equation $f_X(x) = \int f_{Y,X}(y, x)\mathrm{d}y$ look like when Y has a discrete distribution with $\Pr(Y = k) = \pi_k$ for $k = 1, \ldots, K$ (and $\sum_{i=1}^K \pi_i = 1$)? It is merely $\sum_{i=1}^K \pi_i f_{X|y=i}(x)$. (This is the denominator in equation (4.10) on page 139, which was the basic step in developing LDA or QDA.)

Now, from the above with a slight abuse of notation, we have $f_Y(y = k) = \pi_k$, and so

$$f_{Y,X}(y, x) = \pi_1 f_{X|y=1}(x) + \pi_2 f_{X|y=2}(x)$$

and so,

$$f_X(x) = \sum_{i=1}^{K} \pi_i f_{X|y=i}(x).$$

Finally, from $f_{Y|x}(y) = f_{Y,X}(y, x)/f_X(x)$ (Bayes theorem), we have

$$\Pr(Y = k|x) = \frac{\pi_k f_{X|Y=k}}{\sum_{i=1}^K \pi_i f_{X|y=i}(x)}.$$

(This is exactly the same as equation (4.10) on page 139, except that I have used a more explicit notation.)
This is the "answer" to this part of the question; however, you should go ahead and substitute the assumed expressions for the probability densities, as in equation (4.12) in the text.

ii. **Now in terms of** $\Pr(Y = k|x)$**, tell what your rule is for classifying a new observation as either** $Y =$ **1 or 2, given** x_0 **for that observation.**
Also, simplify the mathematical expression on which you base your decision.

The rule that we use in LDA or QDA is to choose the k that corresponds to the largest value of $\Pr(Y = k|x)$ for the given x_0.
We'll use the same rule here.
Note, as in the usual LDA, the only thing we have to look at is the numerator,

$$\frac{\pi_k}{2\theta_k} \mathrm{e}^{-|x_0-\mu_k|/\theta_k}.$$

After simplifying it (by taking logs) the rule is take the k for which

$$\log(\pi_k) - \log(2\theta_k) - |x_0 - \mu_k|/\theta_k$$

is largest.

iii. **Now assume** $\theta_1 = \theta_2 = \theta$**. Tell what your rule is for classifying a new observation as either** $Y =$ **1 or 2, given** x_0 **for that observation.**

This just results in a simplification of the rule in the previous question. The rule now is to choose the k for which

$$\log(\pi_k) - |x_0 - \mu_k|$$

is largest.
In the case of no prior information about the relative frequences or for other reasons, we may just assume that all of the π_ks are equal. In that case, the rule is to choose the k for which

$$|x_0 - \mu_k|$$

is smallest. ("Smallest" makes more sense than "largest" of the negative quantity.)