STA 325: Quiz 3

Total: 56 points

1. [**16 points**] Mark each statement below as TRUE or FALSE. Briefly justify why in 1-2 sentences.

   (a) If one can find a single separating hyperplane for binary data, one can find an infinite number of distinct separating hyperplanes.

   (b) The maximal margin classifier may give a positive training misclassification rate.

   (c) A classifier corresponding to a separating hyperplane is likely to have high variance.

   (d) A classifier with perfect (i.e., 100%) sensitivity or perfect specificity must be a good classifier.

For the next two questions, consider the support vector classification problem:

$$\max_{\beta_0,\beta_1,\cdots,\beta_p,\epsilon_1,\cdots,\epsilon_n} M \quad \text{s.t.} \quad \begin{aligned}[t] &\sum_{j=1}^{p} \beta_j^2 = 1, \\ &y_i(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}) \geq M(1-\epsilon_i), \\ &\epsilon_i \geq 0, \quad C\sum_{i=1}^{n} \epsilon_i \leq 1. \end{aligned}$$

NOTE: this is the same optimization problem in lecture, but with a slight reparametrization of the tuning parameter C.

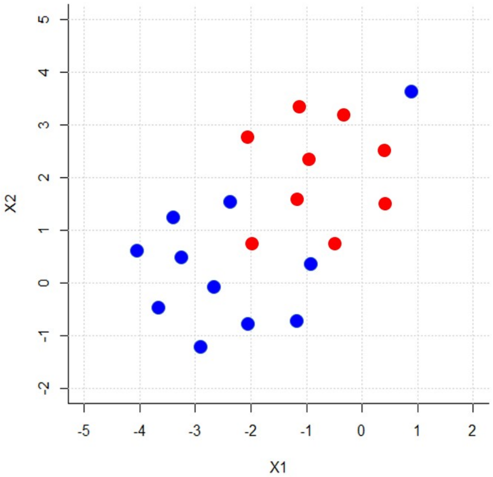
(e) A large choice of C results in a classifier with low variance but high bias.

(f) A small choice of C results in a classifier with wide margins and many support vectors.

(g) As the number of features grows large, support vector machines can be computationally expensive even for a dataset with few points.

(h) With $J = 2$ categories, the baseline-category logit model reduces to a standard logistic regression model.

2. **[11 points]** We have seen that in $p = 2$ dimensions, a linear decision boundary takes the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$.



In the above dataset, clearly such a linear boundary would be insufficient. We thus investigate a few nonlinear decision boundaries below.

(a) **[2 points]** Sketch the curve on the above figure:

$$(1 + X_1)^2 + (2 - X_2)^2 = 4.$$

What is its shape?

(b) **[2 points]** Suppose classifier A uses the above boundary for classification. On the same figure, indicate (annotate on the figure) the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 > 4,$$

as well as the set of points for which

$$(1 + X_1)^2 + (2 - X_2)^2 < 4.$$

Which region should be classified as blue, and which should be classified as red?

(c) [**2 point**] From the above visualized training data, what is the training misclassification error for classifier A? Do you expect its test misclassification error to be lower or higher? Briefly explain.

(d) [**3 points**] Argue that the decision boundary in (c) is linear in terms of X_1, X_1^2, X_2 and X_2^2. Write down the optimization problem for the corresponding support vector classifier (SVC) for an arbitrary cost C.

(e) [**2 points**] Suppose that, despite wanting a nonlinear classifier, we do not know prior information on which nonlinear features to use for SVM training. Write down a reasonable optimization problem to solve in this setting, and briefly explain why this would be appropriate.

3. [**12 points**] Let x be a predictor, e.g., age, and let $Y(x) \in \{1, \cdots, J = 4\}$ be a corresponding ordinal response variable, e.g., stages of cancer. Consider the cumulative logit proportional-odds model:

$$\text{logit}[\mathbb{P}(Y(x) \le j)] = \log\left(\frac{\mathbb{P}(Y(x) \le j)}{1 - \mathbb{P}(Y(x) \le j)}\right) = \alpha_j - \beta x, \quad j = 1, \cdots, J-1. \tag{1}$$

(a) [**2 points**] Recall that the *odds* of an event A is defined as $\mathbb{P}(A)/[1 - \mathbb{P}(A)]$: it quantifies how likely an event is to happen than not. Give an expression for the odds of the event $\{Y(x) \le j\}$ under model (1), for $j = 1, \cdots, J-1$.

(b) [**2 points**] Suppose there are two new patients (patient 1 and 2) with ages x_1 and x_2, respectively. Under model (1), what is the *odds ratio* of patient 2 over patient 1 for the same category j?

(c) [**3 points**] Does the odds ratio in (b) change for different categories j? Interpret what this means for the cancer application.

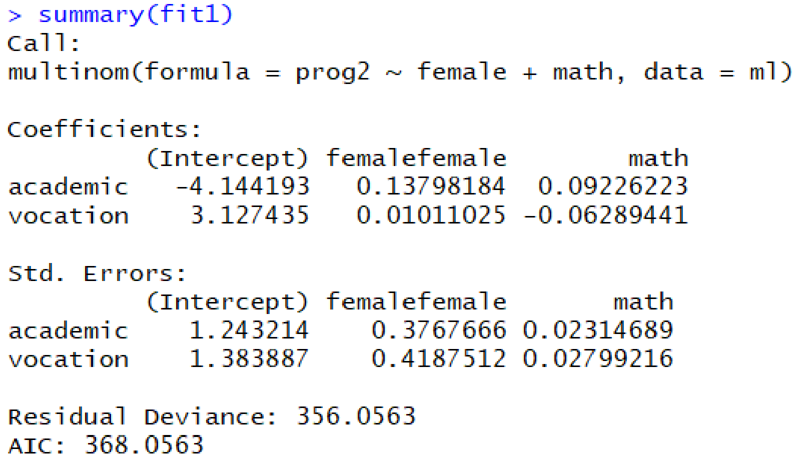
(d) **[2 points]** Consider now a more general cumulative model:

$$\operatorname{logit}[\mathbb{P}(Y(x) \leq j)] = \log\left(\frac{\mathbb{P}(Y(x) \leq j)}{1 - \mathbb{P}(Y(x) \leq j)}\right) = \alpha_j - \beta_j x, \quad j = 1, \cdots, J-1, \tag{2}$$

where a different slope parameter β_j is used for each category j. What is the odds ratio of patient 2 over patient 1, under model (2) for the same category j?

(e) **[3 points]** Does the odds ratio in (d) have the proportional odds property? Interpret what this means for the cancer application.

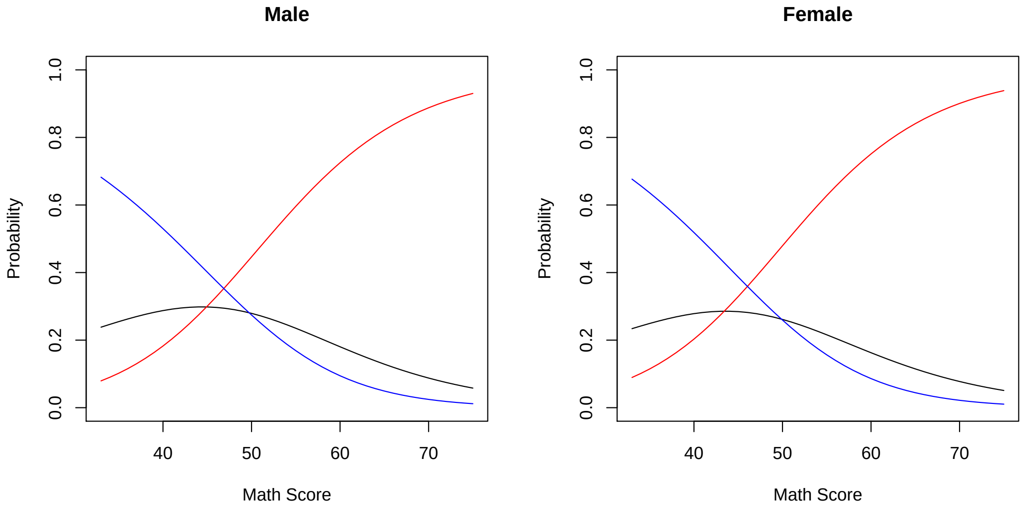
4. [**17 points**] Let's dig deeper into the educational data from HW5 on student program choice. Suppose we have two predictors now, `female` and `math`. The first is a binary predictor with value 1 if the student is female, and 0 if the student is male; the second is the student's score on a math test. Consider the following baseline-category logit model fit for the nominal response variable `prog2` (which takes levels "General" [baseline], "Academic" and "Vocation"):



```
> summary(fit1)
Call:
multinom(formula = prog2 ~ female + math, data = ml)

Coefficients:
         (Intercept) femalefemale        math
academic   -4.144193   0.13798184  0.09226223
vocation    3.127435   0.01011025 -0.06289441

Std. Errors:
         (Intercept) femalefemale       math
academic    1.243214    0.3767666 0.02314689
vocation    1.383887    0.4187512 0.02799216

Residual Deviance: 356.0563
AIC: 368.0563
```

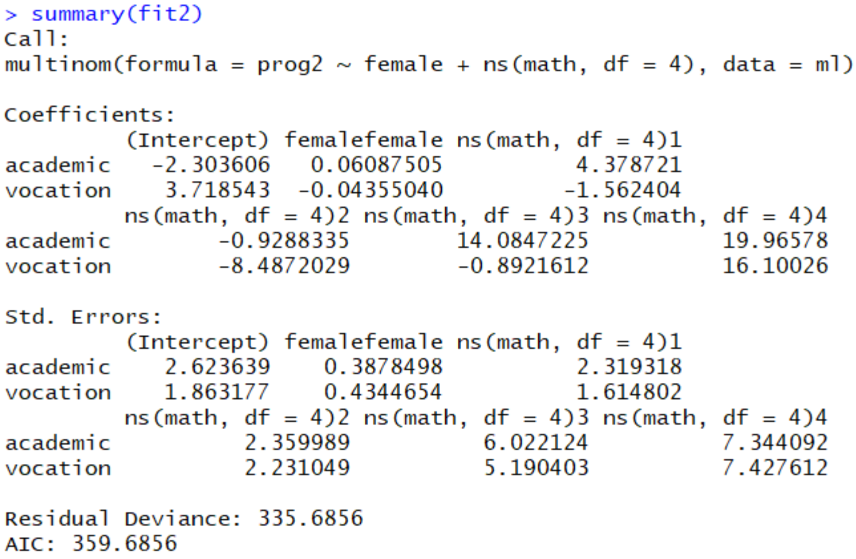
(a) [**3 points**] Let $\pi_G(\texttt{female}, \texttt{math})$, $\pi_A(\texttt{female}, \texttt{math})$ and $\pi_V(\texttt{female}, \texttt{math})$ be the respective class probabilities given predictors `female` and `math`. Write down the modeling equations in terms of the logit probabilities (round estimates to two decimal places). How many equations are needed?

(b) [**3 points**] For each response category ("General", "Academic", "Vocation"), give a formula for the fitted class probabilities (e.g., $\mathbb{P}[Y(x) = \text{General}]$).



(c) **[4 points]** Plotted above are the fitted class probabilities over `math`, for male and female students (black = “General”, red = “Academic”, blue = “Vocation”). Interpret these probabilities in terms of gender and math score differences.

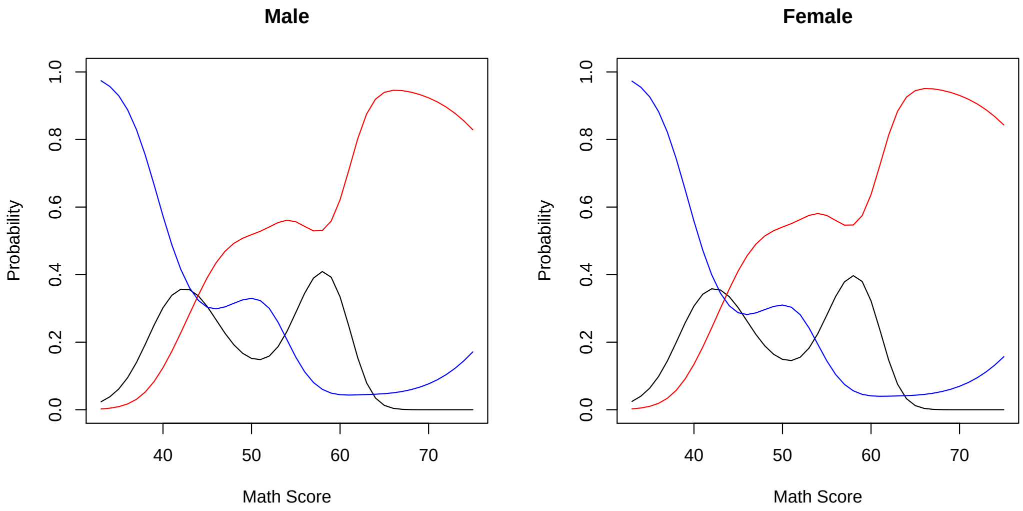
(d) **[4 points]** One potential disadvantage of the model in (a) is that it’s too simplistic, since it only accounts for linear effects in predictors. Suppose we fit the following model with nonlinear effects:



```
> summary(fit2)
Call:
multinom(formula = prog2 ~ female + ns(math, df = 4), data = ml)

Coefficients:
         (Intercept) femalefemale ns(math, df = 4)1
academic   -2.303606   0.06087505          4.378721
vocation    3.718543  -0.04355040         -1.562404
         ns(math, df = 4)2 ns(math, df = 4)3 ns(math, df = 4)4
academic        -0.9288335        14.0847225          19.96578
vocation        -8.4872029        -0.8921612          16.10026

Std. Errors:
         (Intercept) femalefemale ns(math, df = 4)1
academic    2.623639    0.3878498          2.319318
vocation    1.863177    0.4344654          1.614802
         ns(math, df = 4)2 ns(math, df = 4)3 ns(math, df = 4)4
academic          2.359989          6.022124          7.344092
vocation          2.231049          5.190403          7.427612

Residual Deviance: 335.6856
AIC: 359.6856
```

Write down the modeling equations in terms of logit probabilities (round to 2 decimal places, but denote the natural spline as a function of `math`, i.e., `ns`(`math`)). Explain why this provides a more flexible model for multiclass probabilities. Does the data give evidence for this more complex model? Explain.

(e) [**3 points**] Plotted below are the fitted class probabilities from this new model, over `math` for male and female students. Identify a key difference between these fitted probabilities and the ones in part (c), and provide a plausible explanation for this phenomenon given the problem at hand.


Male
Female
Probability
Math Score