Ken Ye
jy294

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.

   True. If one can find a separating hyperplane, it means the data is linearly separable. One can change the direction/position of this hyperplane and find an infinite # of distinct separating hyperplanes, because there always exist some margin allowing for trivial changes of the hyperplane.

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

   False. The MMC assumes a perfect linear seperation of the training data. Therefore, if we were able to fit a MMC, its training misclassification rate must be zero.

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

   True. A seperating hyperplane (e.g. one found by MMC) will have zero training error, however, it's very prone to be affected by the changes in points near the margin. It thus has high variance and could overfit the data.

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

   False. A classifier's sensitivity and specifity need to be balanced and customized for specific purposes. Perfect sensitivity can be achieved by labeling everything as positive, and perfect specificity can be achieved by labeling everything as negative, which are not useful predictions at all, and the test errors could be very high.

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

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

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.

False. A large C leads to smaller ϵ_i's (lower tolerance budget for margin violation) → fewer pts. allowed to violate margin → smaller margin → fewer support vectors determining hyperplane → high variance, low bias. (sensitive to new training pts.)

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

True. A small C in this case means higher tolerance budget for margin violation → larger ϵ_i's → more pts. allowed to violate margin → larger margin → more support vectors determining hyperplane → low variance, high bias. (less sensitive to new training pts.)

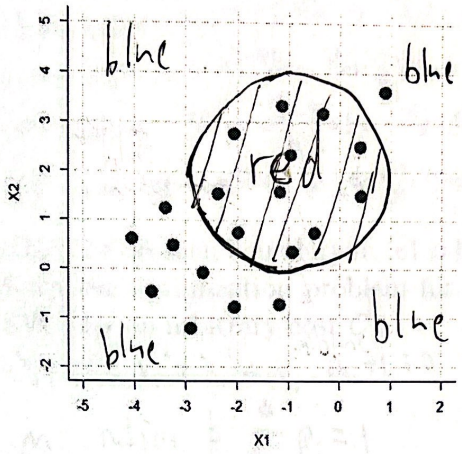
(g) As the number of features grows large, support vector machines can be computationally expensive even for a dataset with few points.

True. With feature expansion, SVM becomes computationally expensive as boundary margin gets more flexible, and more terms are involved in optimization. With p variables, and d^{th} order polynomial, there are $O(d^p)$ features to optimize, which is very computationally expensive as d and p increase.

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

True. The baseline-category logit model has J categories and J-1 equations for modeling. When J=2, there is only 1 equation, which is the equivalent of a standard logistic regression model.

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



classification
/// red
rest blue

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?

It's a circle with radius = 2, centered at (-1, 2).

(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?

Region inside circle → classify red.

Region outside circle → classify blue.

(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.

The training misclassification error rate is $2/20 = 0.1$ (two misclassified blue pts. within circle, where the classification is red). I expect the test error to be higher because the complexity of the classifier (linear might be enough) means high variance (prone to overfitting), and it may not perform well on unseen data.

(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.

$f(x) = (1+x_1)^2 + (2-x_2)^2 - 4 = x_1^2 + x_2^2 + 2x_1 - 4x_2 + 1 = 0 \Rightarrow$ equation for hyperplane

maximize M subject to $\sum_{j=1}^{4} \beta_j = 1$

$\beta_0, \ldots \beta_4, \xi_1, \ldots \xi_n$

$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2) \geq M(1-\xi_i), \quad \xi_i \geq 0, \quad \sum_{i=1}^{n} \xi_i \leq C$

(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.

Since we have not prior info on which nonlinear features to use, using inner-products (kernel functions) can be helpful. In this case, the optimization problem becomes:

$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle$

could be $K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j})^d$, which computes inner-products needed for d dimensional polynomials

$\alpha_i = 0$ for most pts.

$\alpha_i > 0$ only for pts. within/on margin

Solution has the form:

$f(x) = \beta_0 + \sum_{i \in S} \hat{\alpha}_i K(x, x_i)$

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:

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

(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 $\{Y(x) \leq j\}$ under model (1), for $j = 1, \cdots, J-1$.

$$\text{odds of } \{Y(x) \leq j\} = \frac{P(Y(x) \leq j)}{1 - P(Y(x) \leq j)} = e^{\alpha_j - \beta x}$$

(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?

$$\frac{\text{odds}_{\text{patient 2}}}{\text{odds}_{\text{patient 1}}} = \frac{\exp\{\alpha_j - \beta x_2\}}{\exp\{\alpha_j - \beta x_1\}} = \exp\{\beta(x_1 - x_2)\}$$

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

No, the odds ratio in (b) is constant across different categories j: odds ratio $= \exp\{\beta(x_1 - x_2)\} \Rightarrow$ independent of j. This implies that the impact of age on the odds of cancer progression remains consistent across different cancer stages. The odds of being in a higher category relative to a lower category increases by a constant factor β for each unit increase in age, irrespective of the cancer stage.

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

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

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?

$$\frac{\text{odds}_{\text{patient 2}}}{\text{odds}_{\text{patient 1}}} = \frac{\exp\{\alpha_j - \beta_j x_2\}}{\exp\{\alpha_j - \beta_j x_1\}} = \exp\{\beta_j(x_1 - x_2)\}$$

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

No, the odds ratio in (d) does not have the proportional odds property because different slope parameters β_j are used for different stages j of cancer. This implies that the impact of age on the odds of cancer progression varies across different cancer stages.

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? $\Rightarrow$ 2 equations (General is baseline)

$\text{logit}(\pi_A(\text{female}, \text{math})) = -4.14 + 0.14\,\text{female} + 0.09\,\text{math}$

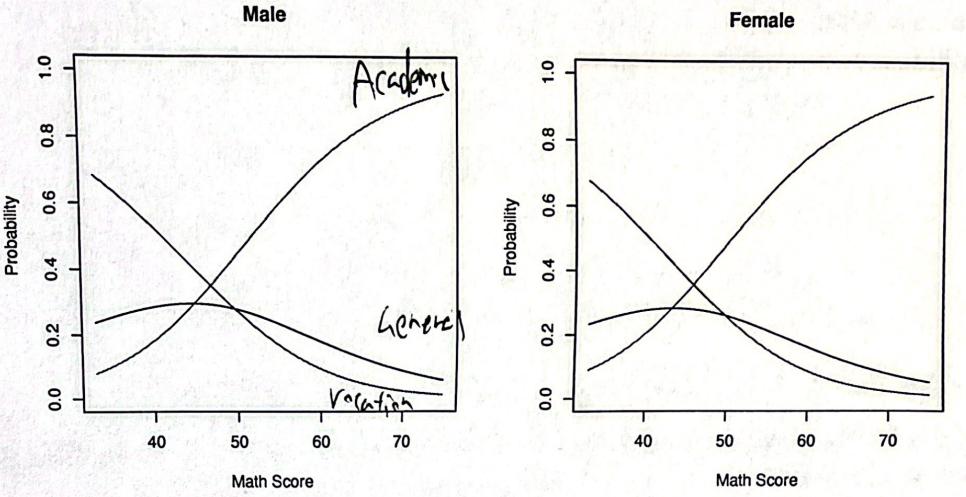
$\text{logit}(\pi_V(\text{female}, \text{math})) = 3.13 + 0.01\,\text{female} - 0.06\,\text{math}$

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

$$\mathbb{P}[Y(x) = \text{General}] = \frac{1}{1 + \exp\{-4.14 + 0.14\,\text{female} + 0.09\,\text{math}\} + \exp\{3.13 + 0.01\,\text{female} - 0.06\,\text{math}\}}$$

$$\mathbb{P}[Y(x) = \text{Academic}] = \frac{\exp\{-4.14 + 0.14\,\text{female} + 0.09\,\text{math}\}}{1 + \exp\{-4.14 + 0.14\,\text{female} + 0.09\,\text{math}\} + \exp\{3.13 + 0.01\,\text{female} - 0.06\,\text{math}\}}$$

$$\mathbb{P}[Y(x) = \text{Vocation}] = \frac{\exp\{3.13 + 0.01\,\text{female} - 0.06\,\text{math}\}}{1 + \exp\{-4.14 + 0.14\,\text{female} + 0.09\,\text{math}\} + \exp\{3.13 + 0.01\,\text{female} - 0.06\,\text{math}\}}$$



(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.

Gender: there isn't observable difference b/w class probs for different genders, meaning gender likely doesn't affect the program outcome by a lot. (Looking at prob distribution, similar)

Math score: as a student's math score increases, their prob of being in an academic program dramatically increases, being in a general program initially increases but then decreases, and being in a vocation program dramatically decreases. This suggests that math score may be a good predictor of a student's program.

(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
```

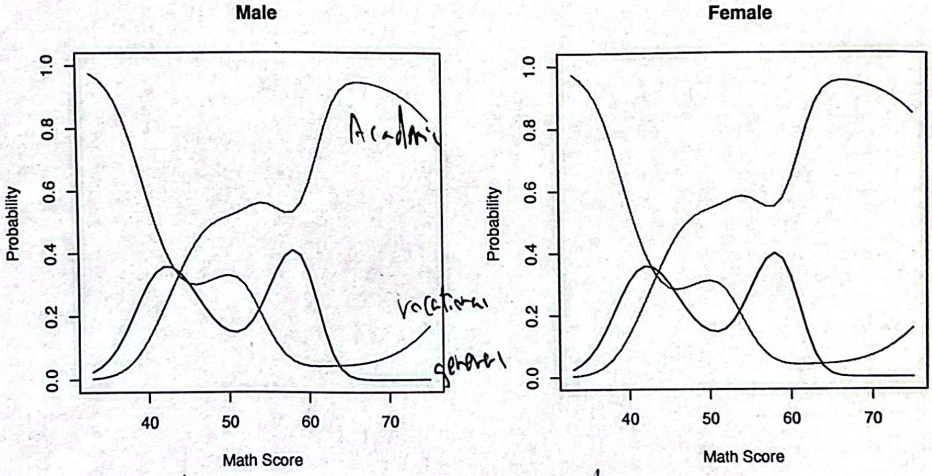
Write down the modeling equations in terms of logit probabilities (round to 2 decimal places, but denote the natural spline as a function of `math`, i.e., `ns(math)`). Explain why this provides a more flexible model for multiclass probabilities. Does the data give evidence for this more complex model?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

$\text{logit}(\pi_A(\text{female}, ns(\text{math}))) = -2.30 + 0.06\,\text{female} + 4.38\,ns(\text{math})$

$\text{logit}(\pi_V(\text{female}, ns(\text{math}))) = 3.72 - 0.04\,\text{female} - 1.56\,ns(\text{math})$

This provides more flexibility because the natural spline allows the relationship b/w math and program to be nonlinear, capturing potential curvature in the relationship (though prone to overfitting). Yes, this more complex model has lower residual deviance (336 < 356) and lower AIC (380 < 568).

(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.



One key difference is how wiggly the pdf lines are for each class in this model compared to the simpler model in part (c). In part (c), the class prob. of Academic monotonely increases as math score increases, and the class prob of vocational monotonely decreases as math score increases; however, with this model, the class probabilities fluctuates, though following similar overall trend. This phenomenon is due to the fact that the natural spline on math captured some nonlinear relationship b/w math score and program, leading to different fitted probs compared to the linear model.

more variability