STA 325: Quiz 3

Total: 66 points + 2 bonus

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 I can have a hyperplane that separates the whole data set perfectly, then I can change the direction or position of this hyperplane for infinite times because there will always exist some margin that allows for trivial changes of position of the hyperplane.

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

False, The MMC assumes that we perfectly separate the dataset, therefore, in terms of the training error we should have the value of zero, in turn we cannot have a positive missclassification rate and high variance because we overfit the data.

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

True, a separating hyperplane means that we separate the data perfectly with zero training error, then we run into the problem of overfitting the data with high variance.

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

False, a classifier with perfect sensitivity or perfect specificity will likly to be affected by the noisy data: for several data points' changes, the classifier may drastically change a lot, therefore it contains the problem of overfitting, high test training error, and high variance although it captures all data points correctly, MMC will also presents incorrect misclassification error.

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

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

huge difference between C in the question, and C in our slide

$$\sum_{i=1}^{n} \varepsilon_i \leq C$$

NOTE: this is the same optimization problem in lecture, but with a 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 choice of C penalizes the classifier with smaller ε_i regularization, the smaller the noisy term, the more complex of our classifer, in turn cause a higher variance with low bias because of the overfitting.

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

True, a smaller C allows for bigger error term ε, so that it allows for misclassified datapoints and a bigger budget of margin violations, and more support vectors may be included within the bigger margin.

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

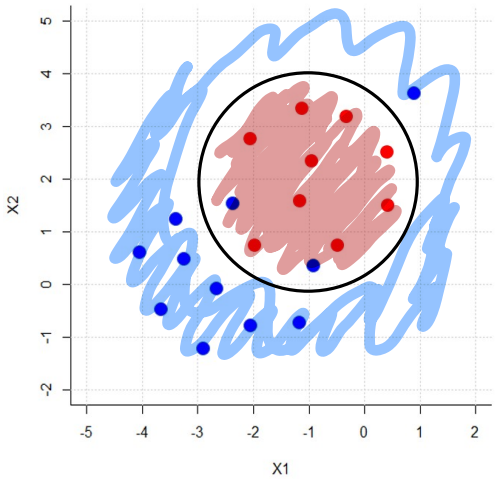
True, with the feature expansion, the SVM is more computationaly expensive because the boundary margin generated this time is more flexible and wiggly as we increase the degree of polynomial classification such as quardatic or cubic function.

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

(as it simplifies to $J-1=1$)

True, the baseline-category with $J=2$ categories will be squashed to a binary logistic regression model with lower dimension, which is the 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 nonlinear decision boundary.

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

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

What is its shape?

It is a circle with radius = 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?

Everything within the circle, which is $(1+X_1)^2+(2-X_2)^2>4$, should be graphed as blue, and everything outside the circle (not on or in the circle), which is $(1+X_1)^2+(2-X_2)^2<4$, 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.

The training misclassfication error is $2/20 = 0.1$ because in total we have two misclassified blue points, and the test error should be higher because this fitting has high variance and low bias and the complexity of this classifier will restrict this model to perform good at bigger population.

(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.

We setup $f(x)=0$,

$f(x) = (1+x_1)^2 + (2-x_2)^2 - 4$

$= x_1^2 + x_2^2 + 2x_1 - 4x_2 + 1$

Therefore, we have $f(x)$ as the linear expression of x_1, x_1^2, x_2, and x_2^2.

We then want to use the Maximal Margin classifier with

$x_1^2 + x_2^2 + 2x_1 - 4x_2 + 1 = 0$

meaning the simplified formula can be

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2$

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

$\beta_0, \ldots, \beta_4$

Then we have

$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \beta_3 x_{i2} + \beta_4 x_{i2}^2) \geq M(1-\varepsilon_i)$

for where $\varepsilon_i \geq 0$ and $\sum_{i=0}^{4} \varepsilon_i \leq C$

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

We can use kernel function if we dont have information about many variables as $f(x) = \beta_0 + \sum_{i \in S} \hat{\alpha}_i K(x_i, x_2)$, because the $\alpha_i = 0$ for most of the points and $\alpha_i > 0$ only for points that are within / on the margin of our classifier, so that we only need $n+1$ parameters.

The kernel implicitiey transform the non-linear features into a single similarity function. $K(x_i, x_{i'}) = \exp\left(-r \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right)$.

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 $j = 1, \cdots, J-1$.

$$\frac{P(Y(x)\leq j)}{1-P(Y(x)\leq j)} = \frac{\exp(\alpha_j - \beta x)}{1+\exp(\alpha_j - \beta x)}, \; j = 1, \cdots, J-1$$

(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?

① odds for patient 1 with age x_1:

$$\frac{P(Y(x_1)\leq j)}{1-P(Y(x_1)\leq j)} = \exp\{\alpha_j - \beta x_1\}$$

② odds for patient 2 with age x_2:

$$\frac{P(Y(x_2)\leq j}{1-P(Y(x_2)\leq j} = \exp\{\alpha_j - \beta x_2\}$$

③ odds ratio $= \dfrac{\text{odds for patient 2}}{\text{odds for patient 1}}$, for j.

$$= \frac{\exp\{\alpha_j - \beta x_2\}}{\exp\{\alpha_j - \beta x_1\}}, \text{ for same } j$$

$$= \exp\{\beta(x_1 - x_2)\}$$

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

The odds ratio doesn't change for different categories j, when we simplifies the equation we have $\exp\{\beta(x_1-x_2)\}$, therefore it has no connection with j.

Meaning that the odds ratio of one patient over another is largely depends on their age and the slope,

For different stages of cancer, the ratio between the patients stay the same, it has the proportional odds property, meaning the slope of which pair of outcomes from cancer development are assumed to be the same segardless of which partition we consider

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

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

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?

odds ration of patient 2 over patient 1 $= \dfrac{\exp\{\alpha_j - \beta_j x_2\}}{\exp\{\alpha_j - \beta_j x_1\}}$

$= \exp\{\beta_j(x_1 - x_2)\}$

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

No, because this time it involves β_j in our odds ratio, so that the odd ratio is different because β is different with $j = 1, \cdots, J-1$.

For each different $j = 1, \cdots, J-1$, we have different odds ratio for the pair of patient 1 and patient 2, which is great for cancer application because we shouldn't suppose that the cancer development rate for two patients are the same all the time regardless of the cancer stages.

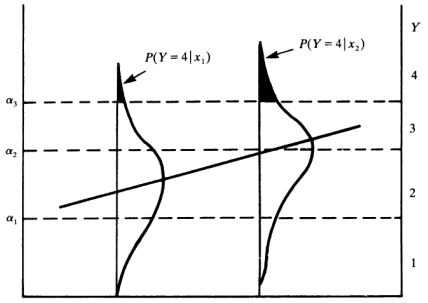
4. **[10 points + 2 bonus]** Suppose the underlying generating mechanism for the ordinal response variable $Y(x)$ follows the two-step procedure. First, for a fixed value of x, simulate the latent (i.e., unobserved) random variable $Z(x)$ from the normal distribution:

$$Z(x) = \mathcal{N}(\beta x, 1).$$

Next, let $Y(x)$ be the following discretization of $Z(x)$:

$$Y(x) = \begin{cases} 1, & \text{if } Z(x) \in (-\infty, \alpha_1] \\ 2, & \text{if } Z(x) \in (\alpha_1, \alpha_2] \\ 3, & \text{if } Z(x) \in (\alpha_2, \alpha_3] \\ \vdots & \\ J, & \text{if } Z(x) \in (\alpha_{J-1}, +\infty) \end{cases}.$$

Another way to view this is that $Y(x)$ is *binned* into the predetermined intervals $(-\infty, \alpha_1]$, $(\alpha_1, \alpha_2]$, ... , $(\alpha_{J-1}, +\infty)$; see below:



$\Phi(x) = P(Z \le x)$

$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left\{\frac{-u^2}{2}\right\} du$

It can be written as many continuous parts of integration $a_1 \cdots a_{J-1}$.

(a) **[3 points]** Show that the cumulative probability $\mathbb{P}[Y(x) \le j] = \Phi(\alpha_j - \beta x)$ for $j = 1, \cdots, J-1$. Here, Φ is the c.d.f. of a standard normal distribution.

$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a_1} \exp\left\{\frac{-u^2}{2}\right\} du + \frac{1}{\sqrt{2\pi}} \int_{a_1}^{a_2} \exp\left\{\frac{-u^2}{2}\right\} \cdot du + \cdots + \frac{1}{\sqrt{2\pi}} \int_{a_{J-1}}^{+\infty} \exp\left\{\frac{-u^2}{2}\right\} \cdot du$ as, $G\{\bar{\pi}_j(x)\} = \alpha_j - \beta x, j = 1 \cdots, J-1$

we can then integrate it as $P(Y(x) \le j) = \Phi(G\{\bar{\pi}_j(x)\}) = \Phi(\alpha_j - \beta x)$ for $j = 1 \cdots, J-1$.

(b) **[3 points]** Using the result in (a), the model can be written as:

$$G\{\mathbb{P}[Y(x) \le j]\} = \alpha_j - \beta x, \quad j = 1, \cdots, J-1. \tag{3}$$

What is the function G? How is this model different from the cumulative logit model in (1)?

The cumulative logit model is $\text{logit}[P(Y \le j)] = \alpha_j + \beta x, j = 1 \cdots, J-1$

For function G, if $\beta > 0$, as x increases, we have smaller probability of falling into lower-valued category, $GP(Y \le j)$ will yield decreasing curve in x, which is completely different with the trends in cumulative logit model in (1).

(c) [**4 points**] What is the odds ratio (see Q3b) under the new model (3)? Does this new model enjoy the proportional odds property? Explain why or why not (a formal proof is not needed, just explain in words).

The odds ratio here is $\frac{P(Y(x) \le j)}{1-(P(Y(x) \le j)} = \frac{\Phi(\alpha_i - \beta x_i)}{1-\Phi(\alpha_i - \beta x_i)}$, $j = 1 \cdots, J-1$

The new model also enjoy the proportional odds property, because the slope will not change for what we partition and G can be considered as tuning parameter which is a control of the model.

(d) [**bonus 2 point**] Give one reason why you might choose this new model (3) over model (1), and one reason why you might choose the earlier model (1) over model (3).

The new model has a smaller variance with little scacrifice on the bias and so I choose the new model over what in (1)

5. **[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? Two equations are needed.

$$Y_i \sim \text{Categorical}(\pi_G(\text{female}, \text{math}), \pi_A(\text{female}, \text{math}), \pi_V(\text{female}, \text{math}))$$

$$\log\left(\frac{\pi_A(\text{female}, \text{math})}{\pi_G(\text{female}, \text{math})}\right) = -4.14 + 0.14\,\text{female}_i + 0.09\,\text{math}_i$$

$$\log\left(\frac{\pi_V(\text{female}, \text{math})}{\pi_G(\text{female}, \text{math})}\right) = 3.13 + 0.01\,\text{female}_i - 0.06\,\text{math}_i$$

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

$$\log\left(\frac{\pi_{ij}}{\pi_{ij^*}}\right) = \beta_{0j} + \beta_j x_i$$

$$\pi_{ij} = \frac{e^{\alpha_j + \beta_j x}}{\sum_h e^{\alpha_h + \beta_h x}}$$

$$P(Y(x) = \text{General}) = \frac{1}{1 + e^{-4.14 + 0.14\,\text{female} + 0.09\,\text{math}_i} + e^{3.13 + 0.01\,\text{female} - 0.06\,\text{math}_i}}$$

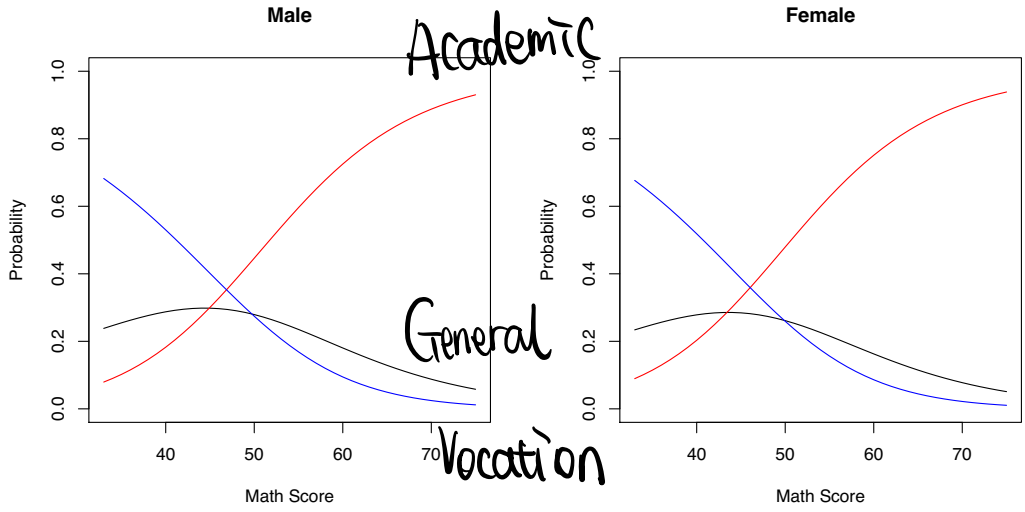
$$\frac{\pi_{A_i}(\text{female}, \text{math})}{\pi_{G_i}(\text{female}, \text{math})} = \exp\{-4.14 + 0.14\,\text{female}_i + 0.09\,\text{math}_i\}$$

$$P(Y(x) = \text{Academic}) = \frac{e^{-4.14 + 0.14\,\text{female}_i + 0.09\,\text{math}_i}}{1 + e^{-4.14 + 0.14\,\text{female} + 0.09\,\text{math}_i} + e^{3.13 + 0.01\,\text{female} - 0.06\,\text{math}_i}}$$

$$\frac{\pi_{V_i}(\text{female}, \text{math})}{\pi_{G_i}(\text{female}, \text{math})} = \exp\{3.13 + 0.01\,\text{female}_i - 0.06\,\text{math}_i\}$$

$$P(Y(x) = \text{Vocation}) = \frac{e^{3.13 + 0.01\,\text{female}_i - 0.06\,\text{math}_i}}{}$$

$1 + e^{4.114 + 0.14\,\text{female} + 0.09\,\text{math}_i} + e^{3.13 + 0.01\,\text{female} - 0.06\,\text{math}_i}$



(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.

For different gender, there is no much differences between each line, meaning that the addition of gender doesn't change the probability that much and there is no interaction term between the predictors. If a student has high math score, it deveases the probability that he/she is in "general" program, increase drastically the probability that he/she is in "Academic program", and greatly decreases the probability that he/she is in "Vocation" 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
```

C

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.

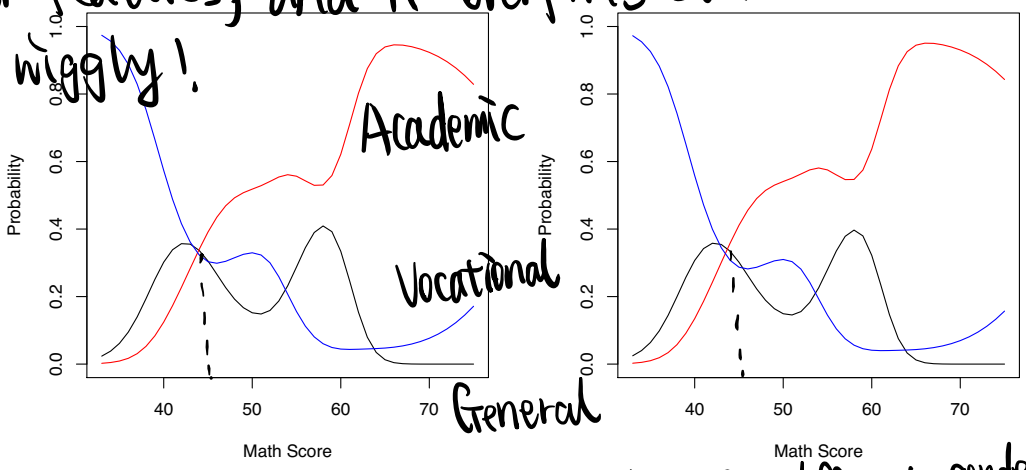
$\text{logit}\left(\frac{\pi_A}{\pi_G}\right) = -2.30 + 0.06\,\text{Isfemale}_i + 4.37\,\text{ns(math)}_i$

$\text{logit}\left(\frac{\pi_V}{\pi_G}\right) = 3.72 - 0.04\,\text{Isfemale}_i - 1.56\,\text{ns(math)}_i$

The residual deviance decrease from 356 to 335, indicating a higher likelihood ratio by the use of natural spline. The AIC also decreases from 368 to 359 and so that this model provides a more flexible model for multiclass probabilities

(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.

The natural spline capture data points and fit the training data better than linear features, and it overfits so that the lines are wiggly!



The probabilities all have the same value of propability at around 45.

Again, there is no changes / differences between lines for different gender as whether the student is male or female doesn't matter compare with the effect of the math score. However, all probability lines don't simply decrease or increase with the increase of Math Score: Math score <40, the probability of student is in "general", "academic" and "vocational" increase, increase, and decrease respectively; For $40<$ math score <60, the probability of "general" decrease then increase, the probability of "vocational" increase and then decrease, the probability of "Academic" increase and then decrease. For $60<$ math score <70, The "general" curve increase then decrease, the "vocational" curve decrease, and the "academic" curve increase a lot. After 70, Academic curve decrease, vocational curve decrease and general curve decrease as well.