Economics 4803/8803: Machine Learning for Economics

Problem Set 3

Due by: **March 29, 5 PM ET**

This problem set is based on topics on classification techniques. You are allowed to work in groups of up to four students, but you must disclose the members of your groups. Individual submissions are required. The code you submit may be identical to the one of the other group members, but I expect comments and answers to the questions to be your own.

Your submission should consist of:

1. a pdf file with responses to the questions in * (do **not** attach any code snippets or question prompts, else you will face a 0.5 point penalty for each)
2. a `R` document with code
3. you should upload these materials separately via the course's Canvas website (please do not email me with your home-work submission).

1 Empirical Problems

This part of the problem set is designed to perform predictive tasks using non-linear models. The task of this problem set is to predict whether a host is a superhost based on host and listing characteristics. Start with the dataset obtained after data cleaning in Problem Set 2. These problems are worth 3 points.

1. Preliminary data cleaning:
   (a) Recode the variable `host_identity_verified` with 't' as 1 and 'f' as 0.
   (b) Remove observations with NA values in the variables `review_scores_rating`, `review_scores_accuracy`, and `review_scores_value`.
2. Analysis:
   (a) Set the seed to 0 and randomly allocate 10% observations to a test set, to be used only in (h).

(b) Estimate a linear probability model with `review_scores_rating` as the predictor.

(c) Repeat (b) using a logit model instead of a linear probability model.

(d) Repeat (b) using a probit model instead of a linear probability model.

(e) Show the coefficients from (b, c, d) in a table. In words, interpret the coefficients of the first three models (b, c, d). How do the estimated relationships compare? Should we interpret these relationships causally? *

(f) Now, add `host_experience` as another covariate, and estimate SVM with a radial kernel for prediction. Set the γ parameter to 0.01 and use 10 fold CV to select the cost parameter from $\{1, 10, 100, 10^3, 10^4\}$.[1]

(g) Obviously, other covariates matter in predicting whether a host is a superhost or not. In addition to `review_scores_rating` and `host_experience`, add `review_scores_accuracy`, `beds`, and `review_scores_value` to the set of predictors. Fit to the data ℓ_1 regularized logistic regression where the full model includes all squared terms and interactions of `review_scores_rating`, `host_experience`, `review_scores_accuracy`, `beds`, and `review_scores_value` (total 20 terms). Using 10-fold CV, find the optimal value of lambda.

(h) Calculate the mean classification error on the test set for each of your 5 models and put them in a table. Explain your results briefly (3-4 lines only).*

2 Conceptual Problems

(Following problems are worth 2 points)

1. Consider the following latent variable model:

$$y_i^* = x_i'\beta + u_i$$

where $u_i|x_i \sim N(0, \sigma^2)$. Define the observed outcome as:

$$y_i = \begin{cases} 1 & \text{if } y_i^* \leq c_1 \\ 2 & \text{if } y_i^* > c_2 \end{cases}$$

[1] Use `tune` and `svm` function in `e1071` library. Refer to Lab 9.6 in ISL textbook for details on implementation. You can use 5 fold CV if the algorithm is slow.

Suppose c_1 and c_2 are known constants.

(a) Derive the conditional log-likelihood function of this model for a random sample $\{y_i, x_i\}_{i=1}^n$

(b) Suppose the data at hand is high-dimensional, i.e. $p \geq n$, where p is the number of covariates. How would you estimate such a model? Write the modified objective function.

2. In lecture 13, we discussed that the log odds ratio for LDA can be expressed as:

$$\ln \frac{\Pr(y=k|x)}{\Pr(y=j|x)} = \ln \frac{\pi_k}{\pi_j} - \frac{1}{2}(\mu_k + \mu_j)'\Sigma^{-1}(\mu_k - \mu_j) + x'\Sigma^{-1}(\mu_k - \mu_j) \tag{1}$$

$$= \alpha_0 + \alpha' x \tag{2}$$

(a) For the simple case with $p = 1$, derive the expression in (1) followed by expressions for α_0 and α_1 in terms of $\pi_k, \pi_j, \mu_k, \mu_j$, and σ^2.

(b) Now, perform the same calculation for QDA, but still take $p = 1$. In this case the log odds will take the form:

$$\ln \frac{\Pr(y=k|x)}{\Pr(y=j|x)} = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{3}$$

Derive the expressions for $\beta_0, \beta_1, \beta_2$ in terms of $\pi_k, \pi_j, \mu_k, \mu_j, \sigma_k^2, \sigma_j^2$.

1.2.e:

- Linear Model:
    - Intercept: The negative intercept (-0.7195582) suggests that when the review_scores_rating is zero, the probability of a host being a superhost is less than zero, which doesn't hold practical significance since the rating scale likely doesn't start at zero.
    - review_scores_rating: The positive coefficient (0.1993090) implies that for each one-point increase in the review score rating, the probability of being a superhost increases by approximately 19.93%. This model implies a direct linear relationship between score rating and the probability of being a superhost.
- Logit Model:
    - Intercept: A very negative intercept (-22.643067) indicates that, in the absence of other factors, the log-odds of being a superhost is quite low.
    - review_scores_rating: The coefficient (4.427234) is much larger than in the LPM, which is typical since it represents the change in log-odds for a one-unit increase in the review score rating. In this case, the odds ratio (exp(4.427234)) would be large, indicating a substantial increase in the odds of being a superhost with each additional point on the review score rating.
- Probit Model:
    - Intercept: Similarly, the intercept is negative (-13.88231), which also indicates low baseline probability.
    - review_scores_rating: The coefficient (2.71555) represents the effect on the probit scale, which can be interpreted as the number of standard deviations change in the latent variable for each one-unit increase in the review score rating.
- The LPM coefficient is less than 1 as it represents a probability change. In contrast, the Logit and Probit models have coefficients that pertain to log-odds and probit scale, respectively, and they are not directly comparable to the LPM coefficient. Generally, coefficients in Logit and Probit models will appear larger than in LPM due to the scales they operate on.
- Logit and Probit models are non-linear and may give different probabilities compared to the LPM for the same value of review_scores_rating, especially when the rating is at the higher or lower ends of the scale.
- The relationships indicated by the models are associative, not necessarily causal. To make a causal claim, we would need to ensure no confounding variables are affecting the relationship, and preferably, the data would come from a randomized controlled experiment. In observational studies like these, it's often safer to interpret the results as associations, noting that higher review scores are associated with a higher probability of being a superhost, without implying that higher scores cause an increase in the probability of being a superhost.

1.2.h:

	Model	Mean_Classification_Error
1	Linear Probability	0.2054521
2	Logit	0.2054521
3	Probit	0.2054521
4	SVM	0.5315824
5	Lasso	0.2034574

We can see that the Linear Probability, Logit, and Probit models all have the same mean classification error, which is somewhat expected as they are different ways of modeling binary outcomes. The SVM (Support Vector Machine) model has a significantly higher error rate, indicating it may not be as well-suited for this particular data as the other models. The Lasso model has the lowest error rate, suggesting it might be the best fit for this data set among the models tested. This could be due to its regularization property that helps in preventing overfitting.

2.1:

(a) The conditional log-likelihood function for this latent variable model, given a sample $\{y_i, x_i\}_{i=1}^n$, is based on the cumulative distribution function of the standard normal distribution, Φ. The likelihood for each observation, depending on whether $y_i = 1$ or $y_i = 2$, will involve Φ and its complement.
The conditional log-likelihood ℓ is the sum of the log-likelihoods for each observation:

$$\ell(\beta, \sigma | \{y_i, x_i\}_{i=1}^n) = \sum_{i=1}^{n} \left[y_i \cdot \log \Phi \left(\frac{c_1 - x_i'\beta}{\sigma} \right) + (1 - y_i) \cdot \log \left(1 - \Phi \left(\frac{c_2 - x_i'\beta}{\sigma} \right) \right) \right]$$

However, since y_i is either 1 or 2 in the observed data, we need to adjust the log-likelihood accordingly:

$$\ell(\beta, \sigma | \{y_i, x_i\}_{i=1}^n) = \sum_{i:y_i=1} \log \Phi \left(\frac{c_1 - x_i'\beta}{\sigma} \right) + \sum_{i:y_i=2} \log \left(1 - \Phi \left(\frac{c_2 - x_i'\beta}{\sigma} \right) \right)$$

(b) In a high-dimensional setting where $p \geq n$, we would typically add a regularization penalty to the log-likelihood function. For example, using L1 regularization (lasso), the modified objective function (also known as the penalized log-likelihood) could be:

$$\mathcal{L}(\beta, \sigma | \{y_i, x_i\}_{i=1}^n, \lambda) = \ell(\beta, \sigma | \{y_i, x_i\}_{i=1}^n) - \lambda \|\beta\|_1$$

where λ is the regularization parameter, and $\|\beta\|_1$ is the L1 norm of β, which is the sum of the absolute values of the coefficients. This encourages sparsity in the coefficient estimates, which is often desirable when $p \geq n$.

2.2:

(a) For LDA, when $p = 1$, we are dealing with a single feature, and the covariance matrix Σ is simply the variance σ^2 in this 1-dimensional case.
Given the log odds ratio formula for LDA:

$$\ln\left(\frac{Pr(y=k|x)}{Pr(y=j|x)}\right) = \ln\left(\frac{\pi_k}{\pi_j}\right) - \frac{1}{2}(\mu_k + \mu_j)'\Sigma^{-1}(\mu_k - \mu_j) + x'\Sigma^{-1}(\mu_k - \mu_j)$$

Since $p = 1$, we can replace x' with x, μ_k' with μ_k, and Σ^{-1} with $1/\sigma^2$, the expression simplifies to:

$$\ln\left(\frac{\pi_k}{\pi_j}\right) - \frac{1}{2\sigma^2}(\mu_k + \mu_j)(\mu_k - \mu_j) + \frac{x}{\sigma^2}(\mu_k - \mu_j)$$

Now, let's expand the quadratic term and simplify:

$$-\frac{1}{2\sigma^2}(\mu_k^2 - \mu_j^2) + \frac{x}{\sigma^2}(\mu_k - \mu_j)$$

$$-\frac{1}{2\sigma^2}(\mu_k^2 - \mu_j^2) = -\frac{\mu_k^2}{2\sigma^2} + \frac{\mu_j^2}{2\sigma^2}$$

So, our expression for the log odds ratio is:

$$\ln\left(\frac{\pi_k}{\pi_j}\right) + \frac{\mu_j^2}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \frac{x}{\sigma^2}(\mu_k - \mu_j)$$

Let's match it with $\alpha_0 + \alpha' x$:
The term not involving x (the intercept term) is α_0:

$$\alpha_0 = \ln\left(\frac{\pi_k}{\pi_j}\right) + \frac{\mu_j^2}{2\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$

And the coefficient of x is α_1:

$$\alpha_1 = \frac{\mu_k - \mu_j}{\sigma^2}$$

(b) For QDA, we must take into account that the covariance matrices for each class k and j may be different. Thus, the term involving the predictor x becomes quadratic and we include σ_k^2 and σ_j^2 instead of Σ.
Given the log odds ratio formula for QDA with $p = 1$:

$$\ln\left(\frac{Pr(y=k|x)}{Pr(y=j|x)}\right) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Since $p = 1$, the expanded QDA expression is not provided in the screenshot, but typically, it would look like this when expanded for a single variable:

$$\ln\left(\frac{\pi_k}{\pi_j}\right) - \frac{1}{2}x^2\left(\frac{1}{\sigma_k^2} - \frac{1}{\sigma_j^2}\right) + x\left(\frac{\mu_k}{\sigma_k^2} - \frac{\mu_j}{\sigma_j^2}\right) - \frac{1}{2}\left(\frac{\mu_k^2}{\sigma_k^2} - \frac{\mu_j^2}{\sigma_j^2}\right)$$

Now we can identify the coefficients β_0, β_1, and β_2 by matching terms:

$$\beta_2 = -\frac{1}{2}\left(\frac{1}{\sigma_k^2} - \frac{1}{\sigma_j^2}\right)$$

$$\beta_1 = \left(\frac{\mu_k}{\sigma_k^2} - \frac{\mu_j}{\sigma_j^2}\right)$$

$$\beta_0 = \ln\left(\frac{\pi_k}{\pi_j}\right) - \frac{1}{2}\left(\frac{\mu_k^2}{\sigma_k^2} - \frac{\mu_j^2}{\sigma_j^2}\right)$$