Midterm exam (125 points)

MATH 6333 Statistical Learning

Opened Thursday 10/28/2021 by 08:00 AM CT

Due Sunday 10/31/2021 by 11:59 PM CT

Problem 1: (25 points)

(a) (15 points) Prove the uniqueness of the predicted value of the lasso regression. That is, show that if $\hat{\beta} \neq \hat{\alpha}$ are two lasso estimates of a multiple linear regression problem with model's matrix X then $\hat{y} = X\hat{\beta} = X\hat{\alpha}$.
(Hint: use the strict convexity of the parabola function

$$(ax + (1-a)y)^2 < ax^2 + (1-a)y^2$$

for $0 < a < 1$ for $x \neq y$ and the convexity of the absolute value (or the triangle inequality)

$$|ax + (1-a)y| \leq a|x| + (1-a)|y|$$

for $0 < a < 1$ for $x \neq y$.)

(b) (10 points) Prove in that case that $\|\hat{\beta}\|_1 = \|\hat{\alpha}\|_1$.

Problem 2: (25 points)

Make a full mathematical description and derivation (as in a proof) of the L_q–regularized logistic regression as a Bayesian estimation problem.

Problem 3: (40 points)

Goal: classify an iris flower based on the inputs: sepal length in cm, sepal width in cm, petal length in cm, and petal width in cm. Iris data and its description are available at https://archive.ics.uci.edu/ml/datasets/iris

Split the data into 80% randomly selected data points to be the training data and the rest 20% to be the testing data. Perform the following list of methods including their cross-validation. Multi-response multiple linear regression.

1) K nearest neighbor
2) Multinomial regression
3) L_1-regularized multinomial regression
4) Elastic-net multinomial regression
5) Linear discriminant analysis
6) Quadratic discriminant analysis
7) Naïve Bayes classification with Gaussian conditional distribution of inputs.

Consider also using higher power terms (e.g. x^2) and interaction terms (e.g. $x_1 x_2$).

Write down a report of the findings of each method and make a table to compare between the different methods. Use confusion matrices as well as testing errors, AIC, and BIC in the comparisons. A conclusion must be made at the end. Please submit the R codes with the report.

Problem 4: (35 points)

Consider a classification problem with general K classes. Assume that the data set is standardized within each class. The goal is to use the naïve Bayes classification method with beta distribution as the conditional distribution of the inputs given the class. You must use the maximum likelihood estimators of the within class parameters.

1) Find $\log\left(\frac{P(G = k|X)}{P(G = K|X)}\right)$ for $k = 1,2, \dots, K-1$.
2) expand the answer of (1) to be in the form of Generalized Additive Model (GAM) and identify the constants a_k, and find the functions $g_{k,i}$,
3) find the discriminant function,
4) find the decision boundary of the classification problem,
5) write down your own function in R of the naïve Bayes classification using the beta distribution for general K. That function shouldn't use any special package for naïve Bayes classification.
6) Apply that function to the iris data in problem 2. Is there any improvement in the classification when compared to results of Problem 2?

——————————————-**End of Exam**——————**Good Luck**——————