2. methodology

In this project, there are 505 predictors, and each predictor has 112167 observes. Not every predictor is important in predicting response. We need find out the important predictors to cut down the computing time. we apply two kinds of methods to finish this project. The first one is used to find the important predictors, including t-test, lasso. The other one is for fit the models to predict whether a person has Asthma or not, including the linear discriminant analysis (LDA), Quadratic discriminant analysis (QDA), Elastic Net.

2.1 Methods for feature selection

**t-test**

t-test can be used to test whether two group data is statistically significant different. The null hypothesis (H0) is that the difference between the two-group means is zero. We compare the p-value get from t-test with the 0.05. If p-value<0.05, reject the null hypothesis. If the values of a predictor responding to healthy samples (Asthma=0) are significantly different from the values of the predictor responding to the disease samples (Asthma=1), the predictor is meaningful to predict Asthma.

**Lasso**



文本

低可信度描述已自动生成(1)

(2)

图示

描述已自动生成

(3)

Formula (1) is the standard linear model. Where Y is a response. X1, X2, X3, … are predictors. *ϵ* is error.

Formula (2) is residual sum of squares (RSS).

Formula (3) the constraint for Lasso. Lasso estimates β0, β1, . . . , βp using the values that minimize

钟表的特写

描述已自动生成

If the coefficient of Xi is zero, Xi is meaningless to predict response.

2.2 Methods for fitting models

**LDA**

LDA estimate probabilities by using Bayes’ Theorem. LDA predicts response by calculating the

文本, 示意图

描述已自动生成

(4)

Where pk(x) represent the probability of response class k when X=x. is prior probability for class k. is the mean, is the standard deviation. By selecting the maximum pk(x), LDA classify the x to class k.

**QDA**

Like LDA, QDA estimate probabilities by using Bayes’ Theorem too. The difference between LDA and QDA is that QDA calculate the covariance matrix for each class k, LDA not. QDA classify the x by selecting the maximum

(5)

Where is a covariance matrix for the kth class k.

**Elastic Net**

Elastic Net estimate the parameters in formular (1) by minimizing

(6)

Where is a parameter to tune, Elastic Net same to ridge when , to lasso when . Elastic Net overcome a shortage of lasso that the predictor selection can be too dependent on data.