CV

2025-12-22

```
data(iris)
set.seed(123)
#CV for evaluation
get_err_logistic = function(x, y, ind_test){
  y_train = y[-ind_test]
  y_test = y[ind_test]
  x_train = x[-ind_test,]
  x_test = x[ind_test,]

  fit = glm(y_train~., family = "binomial", data = x_train)
  coef = fit$coef
  pred_prob = predict(fit, x_test, "response")
  pred_class = (pred_prob > 1/2)
  mean(pred_class != y_test)
  }
cv.logistic = function(x, y, k = 5) {
  ind_list = split(sample(1:length(y)), 1:k)
  score = rep(0, length(ind_list))
  for(i in 1:length(ind_list)){
      score[i] = get_err_logistic(x, y, ind_list[[i]])
  }
  mean(score)
}

y =(iris$Species =="versicolor")
x = iris[,-5]
cv.logistic(x, y, k=5)
```

```
## [1] 0.2866667
```

Take home exercise:

1. sometimes, when the number of parameters becomes too large, (p > n) estimation becomes unfeasible. This is because the matrix X^TX is no longer full rank. One technique that we use to resolve this issue is call Ridge regression. The idea is to add a small lambda on the diagonal of X^TX so that $(X^TX + \lambda * I)$ becomes invertible. Show that for linear regression, the constraint optimization problem

$$\arg\min_{\beta}(y - X\beta)^T(y - X\beta) \quad s.t. \quad \sum_{i=1}^{p} \beta_i \leq t$$

   has solution $\hat{\beta} = (X^TX + \lambda * I)^{-1}X^Ty$. You can refer to page 62-64 of ESL.

2. One problem of this solution is that lambda is not known. Another use of CV is to select the value of lambda. Use the package glmnet to implement Ridge for logistic regression. Use the function cv.glmnet

to select lambda, and glmnet to fit the new model. You can refer to https://cran.r-project.org/web/packages/glmnet/glmnet.pdf. After implementation, apply the method on GALA dataset. Compare the result with logistic regression (with no penalty).