ST340 Lab 6: Validation and the curse of dimensionality

2019–20

Validation

The dataset `SmokeCancer.csv` shows lung cancer rates by U.S. state in 2010, with a number of covariates such as Federal Year 2010 cigarette sales per 100,000.

(a) Read the data file on lung cancer and create a data frame with variables of interest.

```
X = read.table("SmokeCancer.csv", header=TRUE,sep=",",row.names=1)
LungCancer = data.frame(CigSalesRate=100000*X[,"FY2010Sales"]/X[,"Pop2010"],
                        X[,c("CigYouthRate","CigAdultRate","LungCancerRate")])
```

(b) Fit a linear model for LungCancerRate (`?lm` for a reminder about `lm`):

```
summary(lm(LungCancerRate~CigSalesRate+CigYouthRate+CigAdultRate,data=LungCancer))
```

(c) Write a function that takes a formula and does LOOCV (leave one out cross validation) with respect to the squared error of the linear model for the given formula. Use it to find a good linear model for `LungCancerRate` in terms of `CigSalesRate`, `CigYouthRate` and `CigAdultRate`. You could also try using transformations of the covariates by adding terms such as `I(CigSalesRate^2)` and `I(CigSalesRate*CigAdultRate)` to your formulae.

(By good, we mean that it is the optimal, in terms of cross-validation error, linear model using some or all of these covariates.)

(d) The Akaike Information criterion (AIC) and Bayesian Information criterion (BIC) are analytic approximations to the validation step. They are (different) ways of quantifying the trade-off between model complexity (in terms of, e.g. the number of parameters) and the fit to the training data (in terms of likelihood), defined as follows:

- Akaike Information criterion (AIC) = $(2 \times \#\text{parameters} - 2 \times \log(\text{likelihood}))$, and
- Bayesian information criterion (BIC) = $(\log(\text{amount of data}) \times \#\text{parameters} - 2 \times \log(\text{likelihood}))$.

Write a function that takes a formula and then calculates AIC and BIC. Use your function to find a *good* linear model for `LungCancerRate`, as in (b).

The curse of dimensionality

Suppose N points are chosen uniformly at random in the D-dimensional hypercube $[0,1]^D$. Consider a smaller hypercube $H = [0,r]^D$ in the "corner" of $[0,1]^D$.

(a) How big does r have to be for there to be approximately one of the N points lying in H?

(b) How big does r have to be for there to be approximately 10 of the N points lying in H?

(c) How big does r have to be for there to be approximately $\frac{N}{2}$ of the N points lying in H?

Check each of your answers by simulation.

Distance functions

(a) Write a function to calculate the ℓ_1 distances between pairs of row vectors in two matrices:

```
distances.l1 <- function(X,W) {
  # YOUR CODE HERE
}
```

(b) Write a similar function to calculate a matrix of pairwise ℓ_2 distances:

```
distances.l2 <- function(X,W) {
  # YOUR CODE HERE
}
```

(c) Write a similar function to calculate the Mahalanobis distance between the row vectors, given a $D \times D$ covariance matrix S:

```
distances.maha <- function(X,W,S) {
  # YOUR CODE HERE
}
```