**COEN 281**

**Assignment 2**

**Group 1**

Anand Goyal (agoyal2@scu.edu)

Sachin Shinde (sshinde@scu.edu)

**5.** Curse of Dimensionality. Ch.4. Problem 4.7.4.

a.

X is uniformly distributed on [0,1] (given).

We are using observations that are within 10% of the range of X closest to that test observation.

Eg: for x=0.6, we are using those which fall in the range [0.55, 0.65].

That is 0.65 – 0.55 = 0.1

That is we are using only approximately 10% of the available observation.

b.

(X1,X2) are uniformly distributed on [0,1] \* [0,1]

We are using observations which are within 10% of X1 and 10% of the range of X2 closest to that observation.

P = 2

(10%)^2 = 1%

c.

p = 100 features

Observations are uniformly distributed.

Hence for each feature we are considering 10% of the observations.

For p = 100,

(10%)^100 which is almost 0%

d.

Consider a point in 1 dimension.

We can use t% of the data to classify that point.

When we increase the dimension to 2, we use (t%)^2 of the data closest to that point.

(t%)^2 < 1%

As it can be seen that as we go on increasing the dimensions (p), percent of observations considered for predictions decreases exponentially. Thus reducing the correctness of our predictions. As p increases, number of training observation decreases near the test point. This Is a drawback of KNN.

e.

When p = 1, the hypercube is a line.

When p = 2, the hypercube is a square (2 dimensions)

When p = 100, the hypercube is a 100 dimensional cube.

The length of each side of the cube will be pth root of the percentage of observations considered for the predictions.

Therefore,

For p = 1, length of the side of the hypercube is 0.1

For p = 2, length of the side of the hypercube is (0.1)^(1/2) = 0.32

For p = 100, length of the side of the hypercube is (0.1)^(1/100)

**6.** Cross-validation. Ch. 5, Problem 8. Use cv.glm() in library(boot). LOOCV stands for Leave One Out Cross-Validation.

library(boot)

a.

set.seed(1)

Y=rnorm(100)

X=rnorm(100)

Y=X-2\*X^2+rnorm(100)

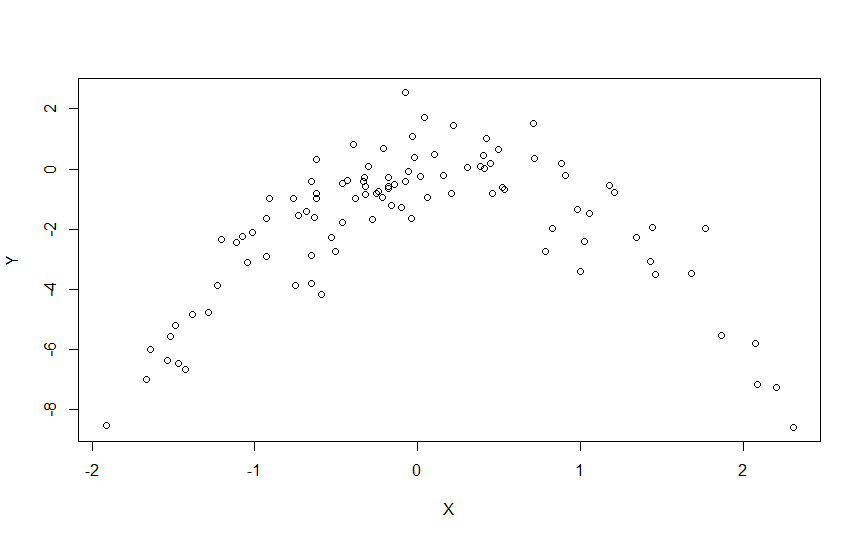
The values are:

n = 100 and p = 2

Model: Y = X - 2X^2 + ε

b.

plot(X,Y)



Observing the plot and the data, we conclude that a curved (quadratic) relationship exists.

c.

set.seed(1)

Data <- data.frame(X, Y)

**i**. Y=β0+β1X+ε

z <- glm(Y ~ X)

cv.glm(Data, z)$delta[1]

[1] 5.890979

**ii**. Y=β0+β1X+β2X^2+ε

z <- glm(Y ~ poly(X,2))

cv.glm(Data, z)$delta[1]

[1] 1.086596

**iii**. Y=β0+β1X+β2X^2+β3X^3+ε

z <- glm(Y ~ poly(X,3))

cv.glm(Data, z)$delta[1]

[1] 1.102585

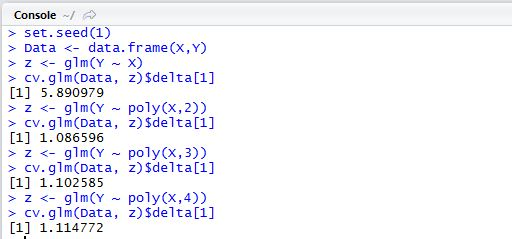
**iv.** Y=β0+β1X+β2X^2+β3X^3+β4X^4+ε

z <- glm(Y ~ poly(X,4))

cv.glm(Data, z)$delta[1]

[1] 1.114772

Output:



**d.**

set.seed(5)

**i**. Y=β0+β1X+ε

z <- glm(Y ~ X)

cv.glm(Data, z)$delta[1]

[1] 5.890979

**ii**. Y=β0+β1X+β2X^2+ε

z <- glm(Y ~ poly(X,2))

cv.glm(Data, z)$delta[1]

[1] 1.086596

**iii**. Y=β0+β1X+β2X^2+β3X^3+ε

z <- glm(Y ~ poly(X,3))

cv.glm(Data, z)$delta[1]

[1] 1.102585

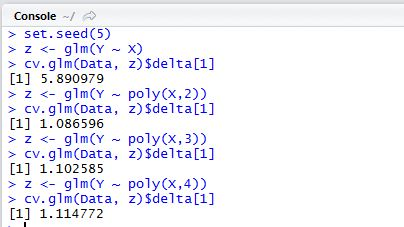
**iv.** Y=β0+β1X+β2X^2+β3X^3+β4X^4+ε

z <- glm(Y ~ poly(X,4))

cv.glm(Data, z)$delta[1]

[1] 1.114772

Output:



Since LOOCV is evaluated ‘n’ times for each observation. The results for question c and d are the same.

e.

As expected, the LOOCV estimation for the MSE(Mean Squared Error) is the minimum for

Y=β0+β1X+β2X^2+ε

This is because the relationship between X and Y is quadratic.

f.

summary(glm(Y ~ X))

Call:

glm(formula = Y ~ X)

Deviance Residuals:

Min 1Q Median 3Q Max

-7.3469 -0.9275 0.8028 1.5608 4.3974

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.8185 0.2364 -7.692 1.14e-11 \*\*\*

X 0.2430 0.2479 0.981 0.329

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 5.580018)

Null deviance: 552.21 on 99 degrees of freedom

Residual deviance: 546.84 on 98 degrees of freedom

AIC: 459.69

Number of Fisher Scoring iterations: 2

summary(glm(Y ~ poly(X,2)))

Call:

glm(formula = Y ~ poly(X, 2))

Deviance Residuals:

Min 1Q Median 3Q Max

-2.89884 -0.53765 0.04135 0.61490 2.73607

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.8277 0.1032 -17.704 <2e-16 \*\*\*

poly(X, 2)1 2.3164 1.0324 2.244 0.0271 \*

poly(X, 2)2 -21.0586 1.0324 -20.399 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 1.06575)

Null deviance: 552.21 on 99 degrees of freedom

Residual deviance: 103.38 on 97 degrees of freedom

AIC: 295.11

Number of Fisher Scoring iterations: 2

summary(glm(Y ~ poly(X,3)))

Call:

glm(formula = Y ~ poly(X, 3))

Deviance Residuals:

Min 1Q Median 3Q Max

-2.87250 -0.53881 0.02862 0.59383 2.74350

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.8277 0.1037 -17.621 <2e-16 \*\*\*

poly(X, 3)1 2.3164 1.0372 2.233 0.0279 \*

poly(X, 3)2 -21.0586 1.0372 -20.302 <2e-16 \*\*\*

poly(X, 3)3 -0.3048 1.0372 -0.294 0.7695

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 1.075883)

Null deviance: 552.21 on 99 degrees of freedom

Residual deviance: 103.28 on 96 degrees of freedom

AIC: 297.02

Number of Fisher Scoring iterations: 2

summary(glm(Y ~ poly(X,4)))

Call:

glm(formula = Y ~ poly(X, 4))

Deviance Residuals:

Min 1Q Median 3Q Max

-2.8914 -0.5244 0.0749 0.5932 2.7796

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.8277 0.1041 -17.549 <2e-16 \*\*\*

poly(X, 4)1 2.3164 1.0415 2.224 0.0285 \*

poly(X, 4)2 -21.0586 1.0415 -20.220 <2e-16 \*\*\*

poly(X, 4)3 -0.3048 1.0415 -0.293 0.7704

poly(X, 4)4 -0.4926 1.0415 -0.473 0.6373

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 1.084654)

Null deviance: 552.21 on 99 degrees of freedom

Residual deviance: 103.04 on 95 degrees of freedom

AIC: 298.78

Number of Fisher Scoring iterations: 2

It is observed that, the linear and quadratic terms are significant. Out of these, the linear term is less significant as compared to the quadratic terms. Cubic and 4th degree terms are not as significant. The reason for this being the relationship between X and Y is quadratic.