**Homework 2**  Sinan DEMİRHAN

**1. Consider the “credit.csv” file where balance (average credit card debt for a number of individuals) is the output attribute. Quantitative input attributes are age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), and rating (credit rating). Qualitative input attributes are gender, student (student status), status (marital status), and ethnicity (Caucasian, African American or Asian).**

**getwd()**

**setwd("C:/Users/Sinan/Desktop/HM2")**

**credit<-read.csv("Credit.csv",header = TRUE)**

**(a) Using linear regression, investigate whether gender has an effect on the credit card balance. What is the average credit card debt for males and females?**

**lineargender<-lm(Balance~Gender,data=credit)**

**summary(lineargender)**

Call:

lm(formula = Balance ~ Gender, data = credit)

Residuals:

Min 1Q Median 3Q Max

-529.54 -455.35 -60.17 334.71 1489.20

Coefficients:

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

(Intercept) 509.80 33.13 15.389 <2e-16 \*\*\*

GenderFemale 19.73 46.05 0.429 0.669

---

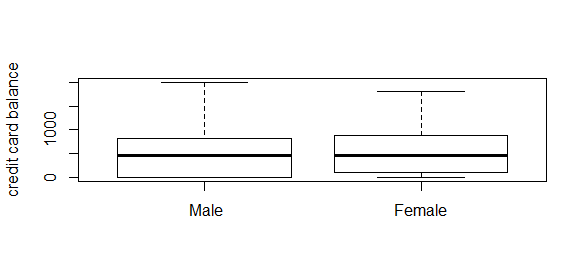
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 460.2 on 398 degrees of freedom

Multiple R-squared: 0.0004611, Adjusted R-squared: -0.00205

F-statistic: 0.1836 on 1 and 398 DF, p-value: 0.6685

**plot(x = credit$Gender, y =credit$Balance)**



**credit[Gender==" Male"]$Balance**

509.8031

**credit[Gender==" Female"]$Balance**

529.5362

If we look at the p value fort he GenderFemale we can see that it is reasonalbe high value and the estimation value for that attribute is very low so we can say that Gender has not any effect on thecredit card balance.

We can see this also from their means.Average credit card debts for males and females are near amounts to each other.

**(b) Using linear regression, investigate whether ethnicity has an effect on the credit card balance.**

**linearethnicity<-lm(Balance~Ethnicity,data=credit)**

**summary(linearethnicity)**

Call:

lm(formula = Balance ~ Ethnicity, data = credit)

Residuals:

Min 1Q Median 3Q Max

-531.00 -457.08 -63.25 339.25 1480.50

Coefficients:

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

(Intercept) 531.00 46.32 11.464 <2e-16 \*\*\*

EthnicityAsian -18.69 65.02 -0.287 0.774

EthnicityCaucasian -12.50 56.68 -0.221 0.826

---

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

Residual standard error: 460.9 on 397 degrees of freedom

Multiple R-squared: 0.0002188, Adjusted R-squared: -0.004818

F-statistic: 0.04344 on 2 and 397 DF, p-value: 0.9575

The situation for ethnicity is the same with the Gender situation.Because the p values are high and the estimation values are low ,we can not say that ethnicity has an effect on the credit card balance.

**(c) We want to investigate whether the effect of income on balance is different for a student compared to a non-student. How can we do this?**

**Student\_credit<-credit[Student=="Yes"]**

**Non\_Student\_credit<-credit[Student=="No"]**

**##We separate the data according to Student attribute.**

**lm\_Y<-lm(Balance~Income,data=Student\_credit)**

**summary(lm\_Y)**

Call:

lm(formula = Balance ~ Income, data = Student\_credit)

Residuals:

Min 1Q Median 3Q Max

-749.55 -372.56 51.38 454.78 730.63

Coefficients:

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

(Intercept) 677.299 118.088 5.736 1.31e-06 \*\*\*

Income 4.219 1.945 2.169 0.0364 \*

---

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

Residual standard error: 468.3 on 38 degrees of freedom

Multiple R-squared: 0.1102, Adjusted R-squared: 0.08674

F-statistic: 4.704 on 1 and 38 DF, p-value: 0.03642

lm\_N<-lm(Balance~Income,data=Non\_Student\_credit)

summary(lm\_N)

Call:

lm(formula = Balance ~ Income, data = Non\_Student\_credit)

Residuals:

Min 1Q Median 3Q Max

-773.4 -325.5 -43.1 314.4 814.0

Coefficients:

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

(Intercept) 200.6232 32.9209 6.094 2.85e-09 \*\*\*

Income 6.2182 0.5784 10.750 < 2e-16 \*\*\*

---

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

Residual standard error: 382.6 on 358 degrees of freedom

Multiple R-squared: 0.244, Adjusted R-squared: 0.2419

F-statistic: 115.6 on 1 and 358 DF, p-value: < 2.2e-16

The students have a effect on the credit card balance but we can easily see that the effect to the credit card balance of non student persons is much more high according to students.

**(d) Perform linear regression first with input attributes “age” and “limit”, and then with “rating” and “limit”. What are your conclusions?**

**linear\_age\_limit<-lm(Balance~Age+Limit,data=credit)**

**summary(linear\_age\_limit)**

Call:

lm(formula = Balance ~ Age + Limit, data = credit)

Residuals:

Min 1Q Median 3Q Max

-696.84 -150.78 -13.01 126.68 755.56

Coefficients:

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

(Intercept) -1.734e+02 4.383e+01 -3.957 9.01e-05 \*\*\*

Age -2.291e+00 6.725e-01 -3.407 0.000723 \*\*\*

Limit 1.734e-01 5.026e-03 34.496 < 2e-16 \*\*\*

---

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

Residual standard error: 230.5 on 397 degrees of freedom

Multiple R-squared: 0.7498, Adjusted R-squared: 0.7486

F-statistic: 595 on 2 and 397 DF, p-value: < 2.2e-16

**linear\_rating\_limit<-lm(Balance~Rating+Limit,data=credit)**

**summary(linear\_rating\_limit)**

Call:

lm(formula = Balance ~ Rating + Limit, data = credit)

Residuals:

Min 1Q Median 3Q Max

-707.8 -135.9 -9.5 124.0 817.6

Coefficients:

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

(Intercept) -377.53680 45.25418 -8.343 1.21e-15 \*\*\*

Rating 2.20167 0.95229 2.312 0.0213 \*

Limit 0.02451 0.06383 0.384 0.7012

---

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

Residual standard error: 232.3 on 397 degrees of freedom

Multiple R-squared: 0.7459, Adjusted R-squared: 0.7447

F-statistic: 582.8 on 2 and 397 DF, p-value: < 2.2e-16

**cor(credit$Limit,credit$Rating)**

0.9968797

**cor(credit$Limit,credit$Age)**

0.1008879

When we look at the age and limit attributes,we can say that they are important for output attribute.However when we fit a linear model with limit and rating attributes these two input attributes does not Show a good effect on the output value because they are highly correlated attributes so we should not include them together into our fitted models.

**2. Using the “UniversalBank.csv” file, split the data set into training (80%) and test (20%) sets with a seed value of “4250”. Apply *k*-fold cross-validation method with *k*=10 on the training set to find the best odd value of *k* (between 1 and 11) to be used in the *k*-NN algorithm. Then report the confusion matrix on the test set using the best *k* value (Use the caret package)**

**getwd()**

**setwd("C:/Users/Sinan/Desktop/dersler/ie 425")**

**bank<-read.csv("UniversalBank.csv",header = TRUE)**

**library(caret)**

**set.seed(4250)**

**trainIndex <- createDataPartition(bank$Personal.Loan, p = .8,list = FALSE,times = 1)**

**train\_bank <- bank[ trainIndex,]**

**test\_bank <- bank[-trainIndex,]**

**anyNA(train\_bank)**

FALSE

**anyNA(test\_bank)**

FALSE

**str(bank)**

'data.frame': 5000 obs. of 14 variables:

$ ID : int 1 2 3 4 5 6 7 8 9 10 ...

$ Age : int 25 45 39 35 35 37 53 50 35 34 ...

$ Experience : int 1 19 15 9 8 13 27 24 10 9 ...

$ Income : int 49 34 11 100 45 29 72 22 81 180 ...

$ ZIP.Code : int 91107 90089 94720 94112 91330 92121 91711 93943 90089 93023 ...

$ Family : int 4 3 1 1 4 4 2 1 3 1 ...

$ CCAvg : num 1.6 1.5 1 2.7 1 0.4 1.5 0.3 0.6 8.9 ...

$ Education : int 1 1 1 2 2 2 2 3 2 3 ...

$ Mortgage : int 0 0 0 0 0 155 0 0 104 0 ...

$ Personal.Loan : int 0 0 0 0 0 0 0 0 0 1 ...

$ Securities.Account: int 1 1 0 0 0 0 0 0 0 0 ...

$ CD.Account : int 0 0 0 0 0 0 0 0 0 0 ...

$ Online : int 0 0 0 0 0 1 1 0 1 0 ...

$ CreditCard : int 0 0 0 0 1 0 0 1 0 0 ...

We should change integer Personal.Loan to factor variable

**train\_bank[["Personal.Loan"]] = factor(train\_bank[["Personal.Loan"]])**

**test\_bank[["Personal.Loan"]] = factor(test\_bank[["Personal.Loan"]])**

**trctrl <- trainControl(method = "cv", number = 10 )**

**fit <- train(Personal.Loan ~ .,method = "knn", tuneGrid = expand.grid(k = 1:11), trControl = trctrl,**

**metric = "Accuracy",** **preProcess = c("center","scale"), data = train\_bank)**

**fit**

k-Nearest Neighbors

4000 samples

13 predictor

2 classes: '0', '1'

Pre-processing: centered (13), scaled (13)

Resampling: Cross-Validated (10 fold)

Summary of sample sizes: 3600, 3600, 3600, 3600, 3600, 3600, ...

Resampling results across tuning parameters:

k Accuracy Kappa

1 0.94650 0.6525720

2 0.94300 0.6233806

3 0.95175 0.6606170

4 0.95250 0.6666393

5 0.95250 0.6568967

6 0.95125 0.6478069

7 0.94950 0.6273790

8 0.95050 0.6344654

9 0.95000 0.6282587

10 0.94850 0.6164031

11 0.94700 0.5987603

Accuracy was used to select the optimal model using the largest value.

The final value used for the model was k = 5.

**preds<-predict(fit,test\_bank)**

**confusionMatrix(preds,test\_bank$Personal.Loan)**

Confusion Matrix and Statistics

Reference

Prediction 0 1

0 890 49

1 9 52

Accuracy : 0.942

95% CI : (0.9257, 0.9557)

No Information Rate : 0.899

P-Value [Acc > NIR] : 8.373e-07

Kappa : 0.6125

Sensitivity : 0.9900

Specificity : 0.5149

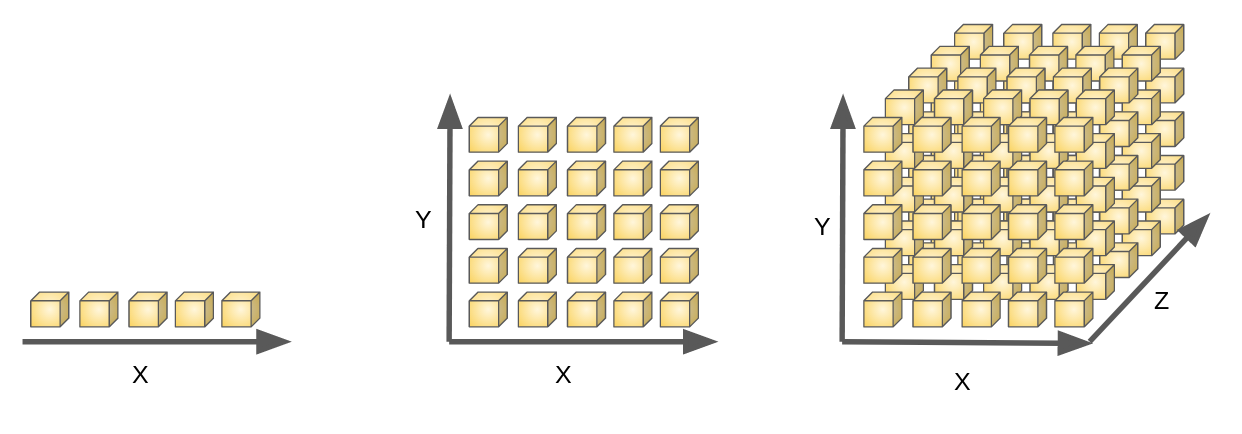
Pos Pred Value : 0.9478

Neg Pred Value : 0.8525

**3. When the number of features p is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the curse of dimensionality, and it ties into the fact that curse of dinon- parametric approaches often perform poorly when p is large. We mensionality will now investigate this curse.**

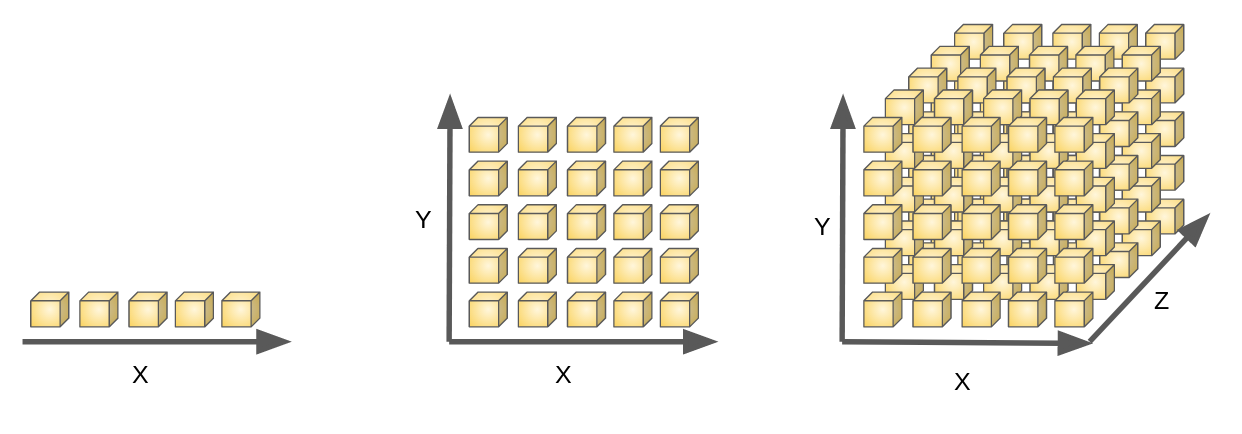
**(a) Suppose that we have a set of observations, each with measurements on p = 1 feature, X. We assume that X is uniformly (evenly) distributed on [0, 1]. Associated with each observation is a response value. Suppose that we wish to predict a test observation’s response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with X = 0.6, we will use observations in the range [0.55, 0.65]. On average, what fraction of the available observations will we use to make the prediction?**

It is approximately 10 % of the observations because the data distributed uniformly.



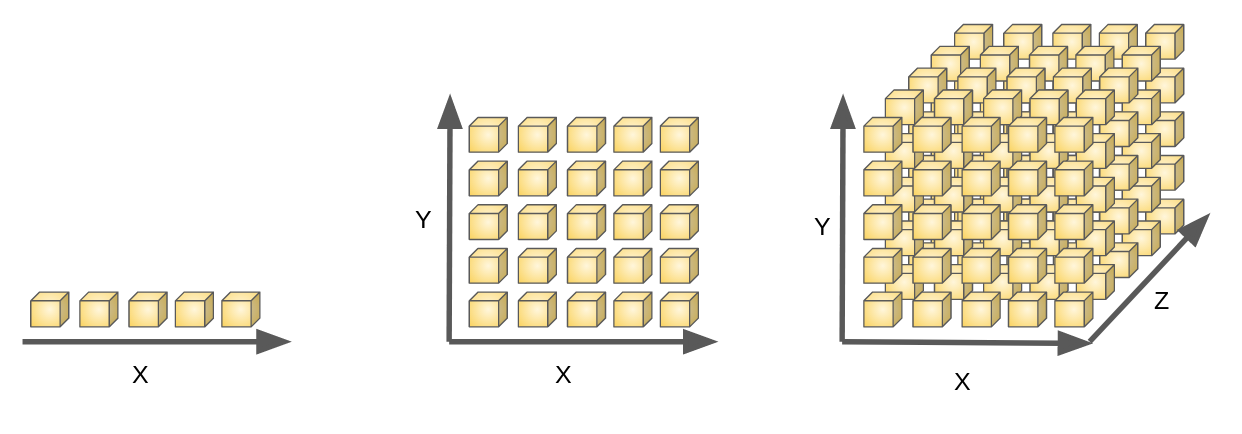
**(b) Now suppose that we have a set of observations, each with measurements on p = 2 features, X1 and X2. We assume that (X1,X2) are uniformly distributed on [0, 1] × [0, 1]. We wish to predict a test observation’s response using only observations that are within 10% of the range of X1 and within 10% of the range of X2 closest to that test observation. For instance, in order to predict the response for a test observation with X1 = 0.6 and X2 = 0.35, we will use observations in the range [0.55, 0.65] for X1 and in the range [0.3, 0.4] for X2. On average, what fraction of the available observations will we use to make the prediction?**

It is approximately 1 % of the observations because of two dimensional



**(c) Now suppose that we have a set of observations on p = 100 features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation’s response using observations within the 10% of each feature’s range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?**

It is approximately 1/(10^100) of the observations because of three dimensional



1/10^p of the observations p is the dimensionality of the observations.It means number of features

**(d) Using your answers to parts (a)–(c), argue that a drawback of KNN when p is large is that there are very few training observations “near” any given test observation.**

When added a feature to a model fraction will be hard for observations so this leads to models to work hard in a much features and small data.To overcome this situation we should increase the amount of data.