HW1

Yunbo Zhu

2/11/2020

Problem 1

1.

```
set.seed(0)

# Define full sample size and simulate feature x
n.full <- 100
x <- runif(n.full,0,10)

# Define linear model as y=1+x+error
y <- 1+0.5*x + rnorm(n.full)

data.full <- data.frame(x=x,y=y)
#plot(x,y)

# Split data
n.test <- 50
index.test <- sample(1:n.full,n.test)
data.train  <- data.full[-index.test,]
data.test  <- data.full[index.test,]

# Fit linear, quadratic and cubic models
linear.train <- lm(y~x,data=data.train)
quad.train <- lm(y~poly(x,degree = 2,raw=T),data=data.train)
cubic.train<- lm(y~poly(x,degree = 3,raw=T),data=data.train)

# Fit linear, quadratic and cubic models
linear.predict.train <- predict(linear.train,newdata=data.train)
quad.predict.train <- predict(quad.train,newdata=data.train)
cubic.predict.train <- predict(cubic.train,newdata=data.train)

# Training error and test error
train.err.linear = mean((data.train$y-linear.predict.train)^2)
train.err.linear
```

```
## [1] 0.8279371
```

```
train.err.quad = mean((data.train$y-quad.predict.train)^2)
train.err.quad
```

```
## [1] 0.8171223
```

```
train.err.cubic = mean((data.train$y-cubic.predict.train)^2)
train.err.cubic
```

```
## [1] 0.8168094
```

2.

```
linear.predict.test <-  predict(linear.train,newdata=data.test)
quad.predict.test <-  predict(quad.train,newdata=data.test)
cubic.predict.test <-  predict(cubic.train,newdata=data.test)

test.err.linear = mean((data.test$y-linear.predict.test)^2)
test.err.quad = mean((data.test$y-quad.predict.test)^2)
test.err.cubic = mean((data.test$y-cubic.predict.test)^2)
test.err.linear
```

```
## [1] 0.731376
```

```
test.err.quad
```

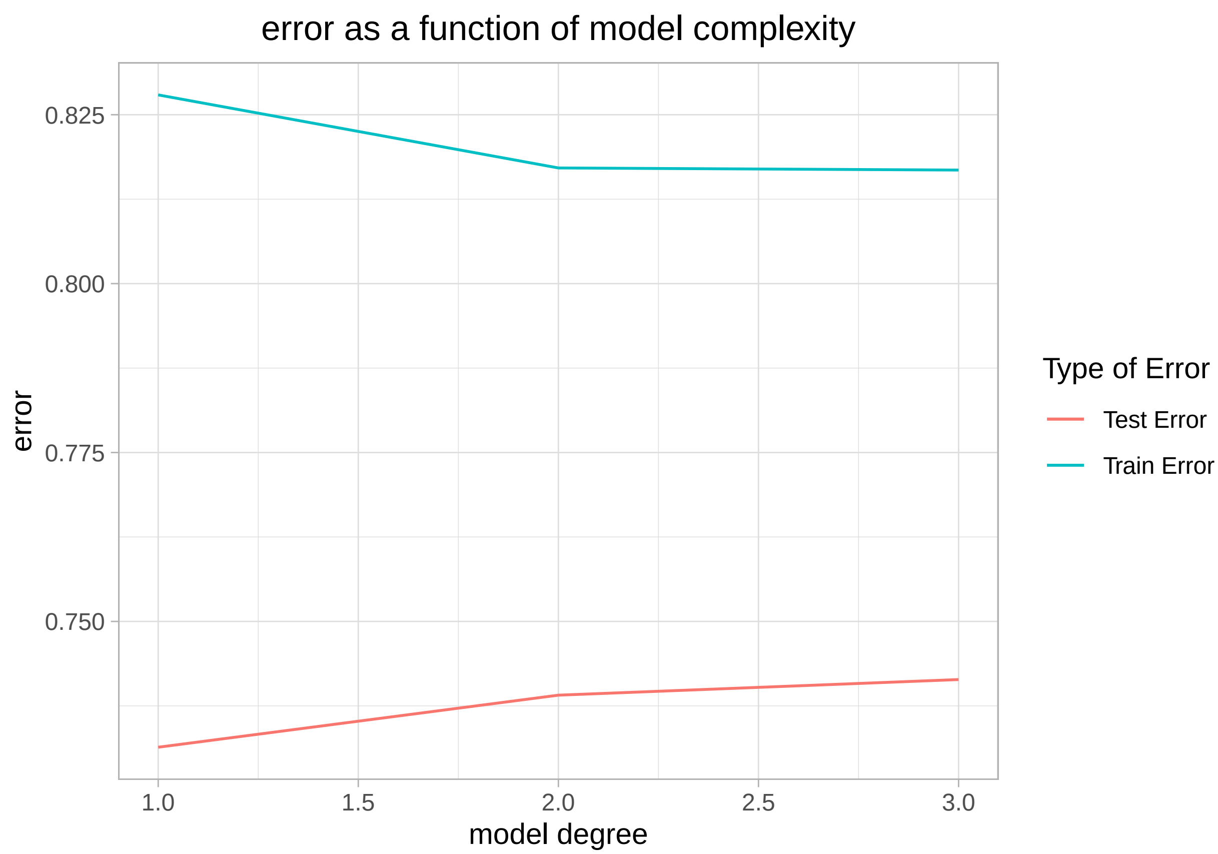
```
## [1] 0.7390816
```

```
test.err.cubic
```

```
## [1] 0.7413943
```

3.

```
model = c(1,2,3)
train.err = c(train.err.linear,train.err.quad,train.err.cubic)
test.err = c(test.err.linear,test.err.quad,test.err.cubic)
library(ggplot2)
err.plot = ggplot()+geom_line(aes(model,train.err,color = 'red'))+geom_line(aes(model,test.err,color ='
  scale_color_discrete(name = "Type of Error" ,labels = c("Test Error","Train Error"))+
  labs(title = "error as a function of model complexity", x = "model degree", y = "error")+
  theme_light()+
  theme(plot.title = element_text(hjust = 0.5))
err.plot
```



Problem 2

1. Maximum Likelihood Estimator: the MLE is the parameter which maximizes the likelihood (joint density) of the data i.e. $\hat{\theta}_{ML} = \arg\max_\theta l(\theta)$, the likelihood function is $l(\theta) = p(x|\theta) = \prod_{i=1}^n p(x_i|\theta)$. Since $\arg\max_y log(f(y)) = \arg\max_y f(y)$, and by logarithm trick $log(\prod_i fi) = \sum_i log(f_i)$, we can get $\hat{\theta}_{ML} = \arg\max_\theta \sum_{i=1}^n p(x_{x_i}|\theta)$. To get the maxima, the parameter must meet the maximality criterion $0 = \sum_{i=1}^n \nabla_\theta logp(x_i|\theta)$

2. derive the ML estimator for μ

$$0 = \sum_{i=1}^n \nabla_\mu ln((\frac{\gamma}{\mu})^\gamma \frac{x^{\gamma-1}}{\Gamma(\gamma)} exp(-\frac{\gamma x}{\mu}))$$
$$0 = \sum_{i=1}^n \nabla_\mu (\gamma ln(\frac{\gamma}{\mu}) + (\gamma - 1)ln(x_i) - ln\Gamma(\gamma) - \frac{\gamma}{\mu}x_i)$$
$$\sum_{i=1}^n(-\frac{\gamma}{\mu} + \frac{\gamma}{\mu^2}x_i) = 0$$
$$-\frac{n\gamma}{\mu} + \frac{\gamma}{\mu^2}\sum_{i=1}^n x_i = 0$$
$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$$

3. instead deriving the formula, show equation to get $\hat{\gamma}$

$$0 = \sum_{i=1}^n \nabla_\gamma ln((\frac{\gamma}{\mu})^\gamma \frac{x^{\gamma-1}}{\Gamma(\gamma)} exp(-\frac{\gamma x}{\mu}))$$
$$0 = \sum_{i=1}^n \nabla_\gamma (\gamma ln(\frac{\gamma}{\mu}) + (\gamma - 1)ln(x_i) - ln\Gamma(\gamma) - \frac{\gamma}{\mu}x_i)$$

$$\sum_{i=1}^{n}(ln(\frac{\gamma}{\mu}) + \gamma * \frac{\mu}{\gamma} * \frac{1}{\mu} + ln(x_i) - \phi(\gamma) - \frac{x_i}{\mu}) = 0$$

$$\sum_{i=1}^{n}(ln(\frac{x_i\hat{\gamma}}{\mu}) + 1 - \phi(\gamma) - \frac{x_i}{\mu}) = 0$$

$$\sum_{i=1}^{n}(ln(\frac{x_i\hat{\gamma}}{\mu}) - (\frac{x_i}{\mu} - 1) - \phi(\gamma)) = 0$$

Problem 3

we want to show that f_0 defined by $f_0 = \arg\max_{y\in[K]} P(y|x)$ minimize $R(f)$

$R(f|x) := \sum_{y\in[K]} L^{0-1}(y, f(x))P(y|x)$ and $R(f) = \int_{R^d} R(f|x)p(x)dx$

if f_0 minimize conditional risks $R(f|x)$, then this $f_0$minimize $R(f)$ as well

$$R(f|x) := \sum_{y\in[K]} L^{0-1}(y, f(x))P(y|x)$$

$$= P(y = f(x)|x) \times \underbrace{\mathrm{L^{0\text{-}1}(f(x), f(x))}}_{0} + \sum_{k\neq f(x)} P(k|x) * \underbrace{\mathrm{L^{0\text{-}1}(f(x), k)}}_{1}$$

$$= \sum_{k\neq f(x)} P(k|x)$$

since $P(k|x)$ sums to 1 over all k, so, $R(f|x) = 1 - P(f(x)|x)$, that is, f_0 maximizes $P(f(x)|x)$, hence minimize $1 - P(f(x)|x)$ at each x.

In summary, $f_0 = \arg\min_{f\in H} R(f|x)$ for all $x \in R^d$, and therefore $f_0 = \arg\min_{f\in H} R(f)$

Problem 4

1. Derive the posterior

$\prod(\theta_1, ..., \theta_K|x_1, ..., x_n) = \dfrac{likelihood * prior}{evidence} = a * \prod_{k=1}^{K} \theta_k^{n_k+\alpha_k-1}$

where a is the normalizing constant

2. Derive the Bayesian MAP

$\hat{\theta}_{MAP} = \arg\max_{\theta} \prod(\theta_1, ..., \theta_K|x_1, ..., x_n)$ where $\theta = (\theta_1, ..., \theta_K)$

drop the normalizing constant, and by logarithm trick, $\hat{\theta}_{MAP} = \arg\max_{\theta} \sum_{k=1}^{K}(n_k + \alpha_k - 1)log(\theta_k)$

by maximizing criterion: $0 = \nabla_{\theta_k} \sum_{k=1}^{K}(n_k + \alpha_k - 1)log(\theta_k)$

by lagrangiam multiplier, we want to maximize function f(x) with constraint g(x) = a, we optimise f(x) - λ(g(x) - a), therefore, in the case of Dirichlet, since $\sum_{k=1}^{K} \theta_k = 1$, thus

$$f(\theta) = log(L(\theta)) = \sum_{k=1}^{K}(n_k + \alpha_k - 1)log(\theta_k)$$

$$g(\theta) = \sum_{k=1}^{K} \theta_k, a = 1$$

$$\text{optimise the lagrangian } \nabla_{\theta_k} \sum_{k=1}^{K}(n_k + \alpha_k - 1)log(\theta_k) - \lambda * (\sum_{k=1}^{K} \theta_k - 1) = 0$$

$$\hat{\theta_k} = \frac{n_k+\alpha_k-1}{\lambda}$$

replace the parameters with their estimates and optimise the lagrangian for lambda,

$$0 = \nabla_\lambda \sum_{k=1}^K (n_k + \alpha_k - 1) log(\frac{n_k+\alpha_k-1}{\lambda}) - \sum_{k=1}^K (n_k + \alpha_k - 1) + \lambda$$
$$\sum_{k=1}^K (n_k + \alpha_k - 1) \times \frac{\lambda}{n_k+\alpha_k-1} \times (-\frac{n_k+\alpha_k-1}{\lambda^2}) + \lambda = 0$$
$$-\sum_{k=1}^K \frac{n_k+\alpha_k-1}{\lambda} + 1 = 0$$
$$\lambda = \sum_{k=1}^K (n_k + \alpha_k - 1)$$
$$\hat{\theta}_{kMAP} = \frac{n_k+\alpha_k-1}{\sum_{k=1}^K (n_k+\alpha_k-1)}$$

3. Derive the "frequintist" MLE of parameters

Similar to above, $\hat{\theta}_{ML} = \arg\max_\theta l(\theta) = \arg\max_\theta \prod_{k=1}^K \theta_k^{n_k}$, therefore to get MLE, $0 = \nabla_{\theta_k} \sum_{k=1}^K n_k log(\theta_k)$ by lagrangian multiplier technique, $\hat{\theta}_{kML} = \frac{n_k}{\sum_{k=1}^K n_k}$