Homework 1

Jack Hart
Collaborators: Arishia Singh, Norman Hong, Shera Ning

Fundamentals and Review

Exercise 1 (Likelihood Estimation)

Problem 1

Let $f(x_i;\theta)$ be the pmf of the geometric distribution:

$$f(X_i;\theta) = (1-\theta)^{(X_i-1)}\theta$$

Suppose we have a random sample of n i.i.d variables from the distribution. The likelihood function can be written as:

$$L(\theta|X_1...X_n) = \prod_{i=1}^{n}(1-\theta)^{(x_i-1)}\theta$$

This can be rewritten as:

$$L(\theta|X_1...X_n) = (1-\theta)^{(\sum_{i=1}^{n}(x_i)-n)}\theta^n$$

Next, derive the log likelihood:

$$l(\theta|X_1...X_n) = (\sum_{i=1}^{n}(x_i)-n)log(1-\theta) + nlog(\theta)$$

Next, differentiate the likelihood:

$$\frac{\partial l}{\partial \theta} = \frac{\sum_{i=1}^{n}(x_i)-n}{1-\theta} + \frac{n}{\theta}$$

To find the MLE we will maximize this likelihood by setting the derivative to 0:

$$\begin{aligned}
0 = -\frac{\sum_{i=1}^{n}(x_i)-n}{1-\theta} + \frac{n}{\theta}\frac{\sum_{i=1}^{n}(x_i)-n}{1-\theta} &= +\frac{n}{\theta} \\
n - n\theta &= \theta\sum_{i=1}^{n}(x_i) - n\theta \\
\hat{\theta}_{MLE} &= \frac{n}{\sum_{i=1}^{n}(x_i)}
\end{aligned}$$

This can be rewritten as:

$$\hat{\theta}_{MLE} = \frac{1}{\frac{\sum_{i=1}^{n}(x_i)}{n}}$$

$$\hat{\theta}_{MLE} = \frac{1}{\bar{x}}$$

Problem 2

Let $f(x_i; a, b)$ be the pmf of the uniform distribution:

$$f(x_i; a, b) = \frac{1}{(b-a)}$$

Suppose we have a random sample of n i.i.d variables from the distribution. From **direct reasoning** we know:

$$a \leq min(X_1, ..., X_n), b \geq max(X_1, ..., X_n)$$

This is because otherwise we wouldn't have been able to sample any observed X_i.

Although trivial, for the sake of rigor, we can take the log likelihood like before and see if it's possible to maximize through it's derivative:

$$L(a, b|X_1...X_n) = \prod_{i=1}^{n} \frac{1}{(b-a)} = \frac{1}{(b-a)^n}$$

$$l(a, b|X_1...X_n) = log(b-x)^{-n} = -nlog(b-a)$$

$$\frac{\partial l}{\partial a} = \frac{n}{b-a}$$
$$\frac{\partial l}{\partial b} = \frac{-n}{b-a}$$

Since $\frac{\partial l}{\partial a}$ is monotonically increasing as n increases, the derviative can be maximized with the smallest a. Converstly, since $\frac{\partial l}{\partial b}$ is monotonically decreasing as n increases, the derviative can be maximized is the largest b.

Therefore, seen in our reasoning above and additional interpretation of the log-loss, the MLEs are as follows:

$$\hat{a}_{MLE} = min(X_1, ..., X_n)$$
$$\hat{b}_{MLE} = max(X_1, ..., X_n)$$

Exercise 2 (Loss Functions)

Problem 1

We know the pdf of this normal distribution to be:

$$f(Y_i; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}}$$

Suppose we have a random sample of n i.i.d variables from the distribution. Therefore the likelihood function, given σ is known, is:

$$L(\mu|Y_i, ..., Y_n) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}}$$

Next we can take the log of this likelihood and algebraically simplify.

$$\begin{aligned} l(\mu|Y_i, ..., Y_n) &= \sum_{i=1}^{n} log((2\pi\sigma^2)^{\frac{-1}{2}}) - \frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2} \\ &= \sum_{i=1}^{n} -\frac{1}{2}log(2\pi\sigma^2) - \frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2} \\ &= -\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \end{aligned}$$

Therefore the negative log likelihood is:

$$-l(\mu|X_i, ..., X_n) = \frac{n}{2}log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

Since σ is known, then the negative log likelihood is the sum of the constant $\frac{n}{2}log(2\pi\sigma^2)$ and another constant term times the L2 loss:

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

Therefore, this is equivalent to L2 loss. Also, the specific term $\sum_{i=1}^{n}(y_i-\mu)^2$ is equivalent to the L2 loss $\sum_{i=1}^{n}(y_i-\hat{y_i})^2$ because **the sum of mean squared differences is the same as the sum of squared point differences.**

Problem 2

We know the pdf of this Laplace distribution to be:

$$f(Y_i;\mu,b) = \frac{1}{2b}e^{-\frac{1}{b}|y_i-\mu|}$$

Suppose we have a random sample of n i.i.d variables from the distribution. Therefore the likelihood function, is:

$$L(\mu|Y_i, ..., Y_n) = \prod_{i=1}^{n}\frac{1}{2b}e^{-\frac{1}{b}|y_i-\mu|}$$

Next we can take the log of this likelihood and algebraically simplify.

$$l(\mu|Y_i, ..., Y_n) = \sum_{i=1}^{n} log(2b^{-1}) - \frac{1}{b}|y_i-\mu| = -nlog(2b) - \frac{1}{b}\sum_{i=1}^{n}|y_i-\mu|$$

Therefore, the negative log-likelihood is:

$$l(\mu|Y_i, ..., Y_n) = nlog(2b) + \frac{1}{b}\sum_{i=1}^{n}|y_i-\mu|$$

The negative log likelihood is the sum of the constant $nlog(2b)$ and another constant term times the L1 loss:

$$\frac{1}{b}\sum_{i=1}^{n}|y_i-\mu|$$

Therefore, this is equivalent to L1 loss. Also, the specific term $\sum_{i=1}^{n}|y_i-\mu|$ is equivalent to the L1 loss $\sum_{i=1}^{n}|y_i-\hat{y_i}|$ because **the sum of mean absolute differences is the same as the sum of absolute point differences.**

Exercise 3 (Decision Rules)

Problem 1

Let $\delta(X)$ be the decision rule/estimator. The accompanying risk function with thus be:

$$R(\mu, \delta(X)) = E(L(\mu, \delta(X))$$

Assume the loss function is the mean squared error and our decision rule is unbiased. Then it can be shown:

$$\begin{aligned} MSE &= Var(X) + Bias(X)^2 \\ &= Var(X) + (E(X) - \delta(X))^2 \\ &= Var(X) + (\mu - \mu)^2 \\ &= Var(X) \end{aligned}$$

So **minimizing the risk is equivalent to minimizing the variance.** Therefore, $R(\mu, \delta(X))$ is minimized by μ when there is no bias. i.e. μ is the optimal decision rule.

Problem 2

Let $\delta(X)$ be the decision rule/estimator. The accompanying risk function with thus be the following for some θ:

$$R(\mu, \delta(X)) = E(L(\theta, \delta(X))$$

Assume our loss function is the mean absolute error. Let $\widetilde{m}$ be the median of some distribution with parameter θ. Now choose a $\hat{\theta}$ that is not the median. Let's consider the losses when we use $\hat{\theta}$ versus $\widetilde{m}$ for *one example*. The loss functions will be:

$$\begin{aligned} L(\hat{\theta}, \theta) &= |\theta - \hat{\theta}| \\ L(\widetilde{m}, \theta) &= |\theta - \widetilde{m}| \end{aligned}$$

The differences of these losses can be written as:

$$L(\widetilde{m}, \theta) - L(\hat{\theta}, \theta) = |\theta - \widetilde{m}| - |\theta - \hat{\theta}| = 2\theta - \widetilde{m} - \hat{\theta}$$

Let's consider two different situations. First when $\widetilde{m} < \theta < \hat{\theta}$ then the following inequality is true:

$$2\theta - \widetilde{m} - \hat{\theta} < 2\hat{\theta} - \widetilde{m} - \hat{\theta} = \hat{\theta} - \widetilde{m}$$

Conversely, when $\widetilde{m} > \theta > \hat{\theta}$ then the following inequality is true:

$$\begin{aligned} 2\theta - \widetilde{m} - \hat{\theta} > 2\hat{\theta} - \widetilde{m} - \hat{\theta} &= \hat{\theta} - \widetilde{m} \\ \widetilde{m} - \hat{\theta} &> 2\theta - \widetilde{m} - \hat{\theta} \end{aligned}$$

Therefore, we can show the following about the differences in these loses:

$$L(\widetilde{m}, \theta) - L(\hat{\theta}, \theta) \leq \begin{cases} \hat{\theta} - \widetilde{m}, & \text{if } \widetilde{m} < \theta \\ \widetilde{m} - \hat{\theta}, & \text{if } \widetilde{m} > \theta \end{cases}$$

Since $\widetilde{m}$ is the median, we know that the conditional density on θ will show the following:

$$\int_{-\infty}^{\widetilde{m}} f(\theta|X)dx = 0.5$$

Therefore the expected value of the differences in the losses can be shown to be 0:

$$\begin{aligned}E[L(\widetilde{m},\theta)-L(\hat{\theta},\theta)] &= 0.5(\hat{\theta}-\widetilde{m})+0.5(\widetilde{m}-\hat{\theta})\\ &= 0.5\hat{\theta}-0.5\hat{\theta}-0.5\widetilde{m}+0.5\widetilde{m}=0\end{aligned}$$

This can be rewritten as:

$$\begin{aligned}E[L(\widetilde{m},\theta)-L(\hat{\theta},\theta)] &\leq 0\\ E[L(\widetilde{m},\theta)]-E[L(\hat{\theta},\theta)] &\leq 0\\ E[L(\widetilde{m},\theta)] &\leq E[L(\hat{\theta},\theta)]\end{aligned}$$

Noted earlier, our risk function is the expected value of the loss. Therefore, this inequality shows that our risk is minimized when we use the median. i.e. **the median is the optimal decision rule**

$$R(\mu,\delta(X))=E(L(\theta,\widetilde{m})$$

Exercise 4 (Convexity)

Problem 1

To prove cross entropy loss is convex with respect to β, we will show that the second derivative is always greater than zero.

First, let's simplify the entire expression algebraically:

$$\begin{aligned}L(y,p) &= -(ylog(p)+(1-y)log(1-p))\\ &= -(ylog(\frac{exp(\beta x)}{1+exp(\beta x)})+(1-y)log(\frac{1}{1+exp(\beta x)}))\\ &= -(y\beta x-ylog(1+exp(\beta x))-log(1+exp(\beta x))+ylog(1+exp(\beta x)))\\ &= -y\beta x+log(1+exp(\beta x)))\end{aligned}$$

Next, we will take the first derivative with respect to β:

$$\frac{\partial l}{\partial \beta}=-yx+\frac{1}{1+exp(\beta x)}exp(\beta)x=-yx+\frac{x}{1+exp(-\beta x)}$$

Next we'll take the second derivative with respect to β (this involves using the quotient rule):

$$\frac{\partial l}{\partial^2 \beta}=\frac{(0)(1+exp(-\beta x))-(x^2)exp(-\beta x))}{(1+exp(-\beta x))^2}=\frac{x^2exp(-\beta x)}{(1+exp(-\beta x))^2}$$

We know that $exp(-\beta x)\geq 0$ and $x^2\geq 0$. Therefore, $x^2exp(-\beta x)\geq 0$ always. And also therefore, $(1+exp(-\beta x))^2\geq 0$ always.

$$\therefore \frac{x^2exp(-\beta x)}{(1+exp(-\beta x))^2}=\frac{\partial l}{\partial^2\beta}\geq 0$$

Since the second derivative is always greater than or equal to 0, the function is convex.

Problem 2

We will prove this function is not convex by taking the second derivative of this function and show that it is **not** always positive.

The function we want to differentiate is the squared error loss:

$$L(y,\beta) = (y - \frac{1}{1+exp(-\beta x)})^2 = (y - \frac{exp(\beta x)}{1+exp(\beta x)})^2$$

First, let's take the first derivative with respect to β (this involves using the quotient rule):

$$\begin{aligned}
\frac{\partial L}{\partial \beta} = 2(y - \frac{exp(\beta x)}{1+exp(\beta x)})(-\frac{e^{\beta x}x\left(1+exp(\beta x)\right) - xexp(x\beta)exp(\beta x)}{\left(1+exp(\beta x)\right)^2}) \\
= 2(y - \frac{exp(\beta x)}{1+exp(\beta x)})(-\frac{xexp(\beta x)}{\left(1+exp(\beta x)\right)^2}) \\
= (2y - \frac{2exp(\beta x)}{1+exp(\beta x)})(-\frac{xexp(\beta x)}{\left(1+exp(\beta x)\right)^2}) \\
= -\frac{2exp(\beta x)x\left(y\left(1+exp(\beta x)\right) - exp(\beta x)\right)}{\left(1+exp(\beta x)\right)^3}
\end{aligned}$$

Next, we'll take the second derivative of this function. Again, this mostly involves applying the quotient rule. This derivative is tedious, so I simplify a lot of the steps together here:

$$\begin{aligned}
\frac{\partial L}{\partial^2 \beta} = -2x\frac{(1+exp(\beta x))^3(2yx*exp(2\beta x) - 2x*exp(2\beta x) + yx*exp(\beta x)))}{\left(\left(1+exp(\beta x)\right)^3\right)^2} - \\
2x\frac{(-(1+exp(\beta x))^2 3x*exp(\beta x)*exp(\beta x)\left(y\left(1+exp(\beta x)\right)exp(\beta x)\right)}{\left(\left(1+exp(\beta x)\right)^3\right)^2} \\
= -\frac{2exp(\beta x)x^2\left(y + exp(2bx) - 2exp(\beta x) - y*exp(2\beta x)\right)}{\left(1+exp(\beta x)\right)^4}
\end{aligned}$$

Let $\beta = 0$. Then this derivative would be:

$$\frac{\partial L}{\partial^2 0} = -\frac{2exp(0)x^2\left(y + exp(0) - 2exp(0) - y*exp(0)\right)}{\left(1+exp(0)\right)^4} = -\frac{2x^2}{16}$$

In this case, $\frac{\partial L}{\partial^2 \beta} < 0$. Thus, the function is not convex.

Exercise 5 (Decision Boundaries)

Problem 1

When $\theta = 0$, then we can find the following functional form of $f_\theta(x)$:

$$f_\theta(x) = \frac{1}{1+exp(-(0+\sum_{i=1}^n 0x_i))} = \frac{1}{1+exp(0)} = \frac{1}{2}$$

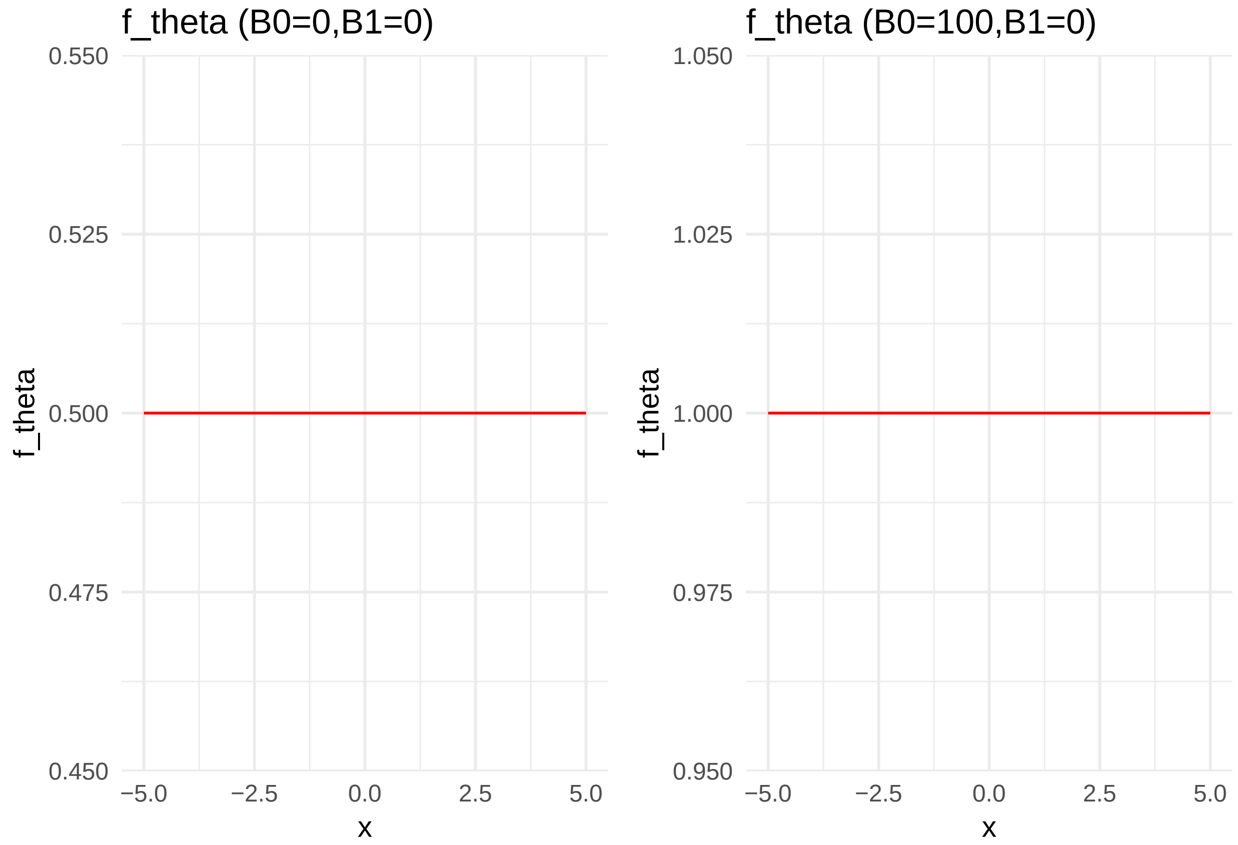
The following code sets up the functions for ploting f_theta for the simple case of n = 1. *This is the basis for how plots are made in problems 1 and 2 of this exercise.*

```r
library(ggplot2)
library(dplyr)
library(gridExtra)

# define f_theta function
f_theta <- function(theta_0, theta_1, x){
  #takes in the thetas and the x_n
  output <- 1 / (1 + exp(-1*(theta_0 + theta_1*x)) )
  return(output)
}

# define a classification function for a matrix of data
classify_f_theta <- function(theta_0, theta_1, X){
  #takes the thetas and a matrix X that contains entire data
  f_output <- c()
  for(examp in 1:nrow(X)){
      f_output <- c(f_output, f_theta(theta_0,theta_1, X[examp,]))
  }
  return(f_output)
}

# define function for logit -- takes the matrix  data
# Proved the logit is equal to the linear combination of theta above
logit_f_theta <- function(theta_0, theta_1, X){
  #takes the thetas and a matrix X that contains entire data
  logit_output <- c()
  for(examp in 1:nrow(X)){
      logit <- theta_0 + theta_1*examp
      logit_output <- c(logit_output, logit)
  }
  return(logit_output)
}
```

Let's plot $f_\theta(x)$ in the case where $\theta = 0$:

This indicates that the probability an example is in class A is 0.5 for all values of x. Therefore, the decision threshold for $f_{\theta=0}(0)$ is **indeterminate** for all values of $f_{\theta=0}(0)$ (assuming our threshold is 0.5 for classification).

If we set $\theta_0 = 100$, the following is a plot of the new logit function. In this case, $f_{\theta=0}(0) \approx 1$ for all values of x. This means that we will **always classify an example as from class A**.



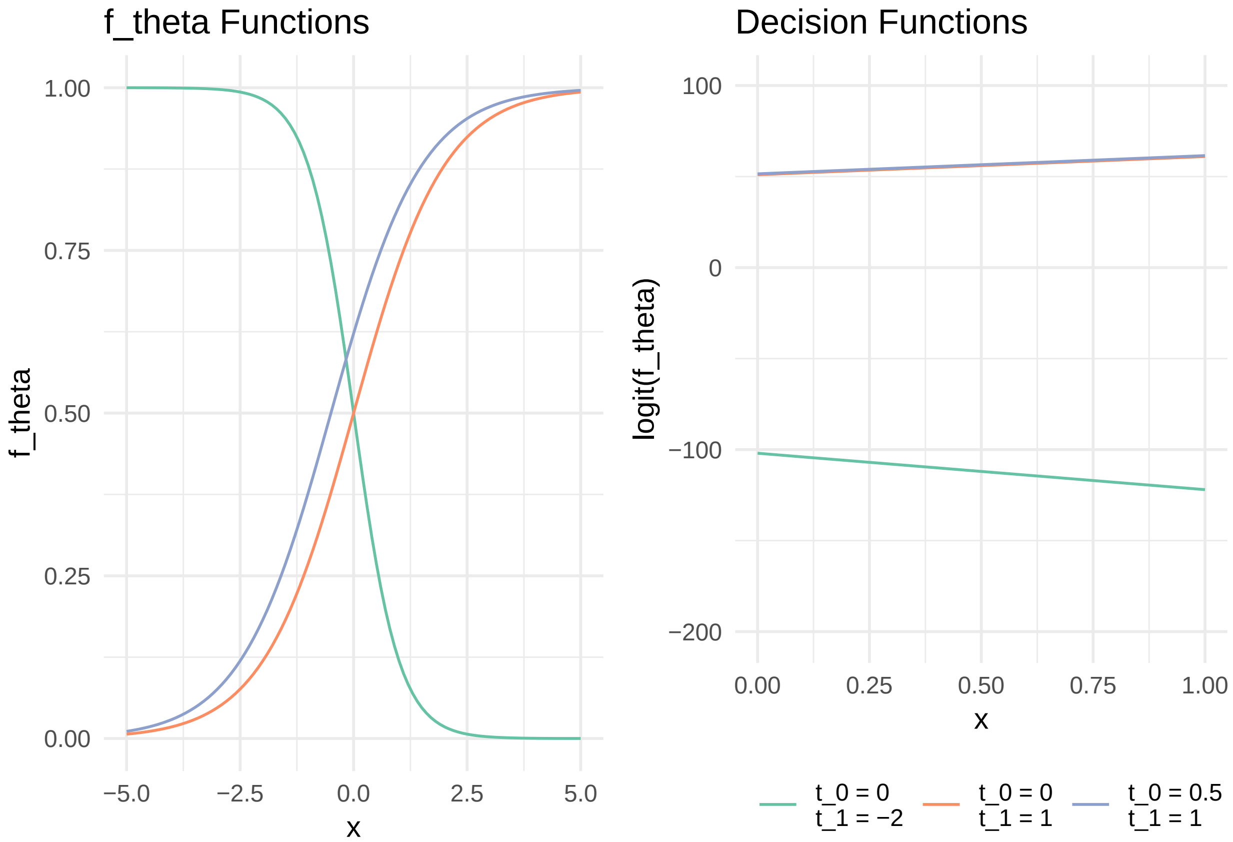
A Note: The code that created this plot can be found in the RMarkdown file.

Problem 2

We can show that $\theta_0 + \theta_1 x$ is the linear separating hyperplane of $f_\theta(x)$ by taking the logit (the link function) of $f_\theta(x)$:

$$\begin{aligned}
logit(f_\theta(x)) &= log(\frac{f_\theta(x)}{1 - f_\theta(x)}) \\
&= log(\frac{\frac{1}{1+exp(-\theta x)}}{1 - \frac{1}{1+exp(-\theta x)}}) \\
&= log(\frac{\frac{1}{1+exp(-\theta x)}}{1 - \frac{exp(-\theta x)+1-1}{1+exp(-\theta x)}}) \\
&= log(\frac{1+exp(-\theta x)}{(1+exp(-\theta x))exp(-\theta x)}) \\
&= log(\frac{1}{exp(-\theta x)}) \\
&= -log(exp(-\theta x)) = -(-\theta x) = \theta x
\end{aligned}$$

Again for the simple case of 1, the following plots $f_\theta(x)$ and it's log-odds (the logit), which shows the linear decision boundary created, for different values of θ. As you can see, when θ_1 is negative, then $f_\theta(x)$ inverts and so does the slope of the decision boundary. Also if θ_0 is not zero, then the intercept of the decision boundary changes and $f_\theta(x)$ becomes a bit less steep.



A Note If we were fitting this logistic model, highly skewed data would skew a model towards usually predicting one class over the other. An example of the intuition behind this was in problem 1, which showed that the predicted probabilities can be heavily skewed if the parameters are set a certain way.

Parametric learning

Exercise 6 (Sufficient Statistic)

We know the pdf of this normal distribution to be:

$$f(X_i; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2})$$

Since each event X_i is i.i.d and σ^2 is known, then the joint pdf can be written as:

$$f_n(X_i, ..., X_n|\mu) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2})$$

The following algebraic manipulation can be made:

$$\begin{aligned}
f_n(X_i, ..., X_n|\mu) &= \frac{1}{\sigma^n (2\pi)^{\frac{2}{n}}} e^{\sum_{i=0}^{n} -\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}} \\
&= \frac{1}{\sigma^n (2\pi)^{\frac{2}{n}}} e^{-\frac{1}{2\sigma^2}\sum_{i=0}^{n}(x_i-\mu)^2} \\
&= \frac{1}{\sigma^n (2\pi)^{\frac{2}{n}}} e^{-\frac{1}{2\sigma^2}\sum_{i=0}^{n}(x_i^2-2\mu x_i+\mu^2)} \\
&= \frac{1}{\sigma^n (2\pi)^{\frac{2}{n}}} e^{-\frac{\sum_{i=0}^{n} x_i^2 - 2\mu \sum_{i=0}^{n} x_i + n\mu^2}{2\sigma^2}} \\
&= \frac{1}{\sigma^n (2\pi)^{\frac{2}{n}}} e^{-\frac{\sum_{i=0}^{n} x_i^2}{2\sigma^2}} e^{\frac{\mu}{\sigma^2}\sum_{i=0}^{n} x_i - \frac{n\mu^2}{2\sigma^2}}
\end{aligned}$$

In this new form we can replace $\sum_{i=0}^{n} x_i$ with $\frac{\sum_{i=0}^{n} x_i}{n}$:

$$f_n(X_i, ..., X_n|\mu) = \frac{1}{\sigma^n (2\pi)^{\frac{2}{n}}} e^{-\frac{\sum_{i=0}^{n} x_i^2}{2\sigma^2}} e^{\frac{\mu n \sum_{i=0}^{n} x_i}{\sigma n} - \frac{n\mu^2}{2\sigma}}$$

Let $T(X) = \frac{\sum_{i=0}^{n} x_i}{n} = \bar{x}$, $h(X) = \frac{1}{\sigma^n (2\pi)^{\frac{2}{n}}} e^{-\frac{\sum_{i=0}^{n} x_i^2}{2\sigma^2}}$, and $g(T(X)|\mu) = e^{\frac{\mu n \bar{x}}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}}$

Now we have written the joint pdf of the sample X as:

$$f(X_i, ..., X_n|\mu) = g(T(X)|\mu)h(X)$$

Therefore, by the **factorization theorem** $\bar{x}$ is a sufficient statistic.

Exercise 7 (Ancilliarity)

Because $F(x-\theta)$ is the cdf of a location parameter family, we know:

$$F(r|\theta) = Pr_\theta(R \leq r)$$

Given $X_i = Z_i + \theta$, we can rewrite this as:

$$\begin{aligned}
F(r|\theta) &= Pr_\theta(max(X_i) - min(X_i) \leq r) \\
&= Pr_\theta(max(Z_i+\theta) - min(Z_i+\theta) \leq r)
\end{aligned}$$

Since θ is a constant, it can be taken out of the min and max functions:

$$\begin{aligned}
F(r|\theta) &= Pr_\theta(max(Z_i) + \theta - min(Z_i) - \theta \leq r) \\
&= Pr_\theta(max(Z_i) - min(Z_i) \leq r)
\end{aligned}$$

Therefore, $Z_i \sim F(z)$ is independent of θ. Thus, the range of this distribution of R is independent of θ.

Exercise 8 (Completeness)

We can write the joint pdf to be:

$$f_n(X_i, ..., X_n|\mu, \mu^2) = \prod_{i=1}^{n} \frac{1}{\mu\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x_i-\mu)^2}{\mu^2}}$$

First we need to find a sufficient statistic using the factorization theorem. To do so, the following algebraic manipulation can be made of the joint pdf:

$$\begin{aligned} f_n(X_i, ..., X_n|\mu, \mu^2) &= \frac{1}{\mu^n (2\pi)^{\frac{2}{n}}} e^{\sum_{i=0}^{n} -\frac{1}{2}\frac{(x_i-\mu)^2}{\mu^2}} \\ &= \frac{1}{\mu^n (2\pi)^{\frac{2}{n}}} e^{\sum_{i=0}^{n} -\frac{(x_i^2-2\mu x_i+\mu^2)}{2\mu^2}} \\ &= \frac{1}{\mu^n (2\pi)^{\frac{2}{n}}} e^{-\frac{\sum_{i=0}^{n} x_i^2}{2\mu^2}+\frac{\sum_{i=0}^{n} x_i}{\mu}-\frac{n}{2}} \\ &= \frac{1}{(2\pi)^{\frac{2}{n}}} \frac{1}{\mu^n} e^{-\frac{\sum_{i=0}^{n} x_i^2}{2\mu^2}+\frac{\sum_{i=0}^{n} x_i}{\mu}-\frac{n}{2}} \end{aligned}$$

Let $T(X) = \{\sum_{i=0}^{n} x_i^2, \sum_{i=0}^{n} x_i\}$,$h(X) = \frac{1}{(2\pi)^{\frac{2}{n}}}$, and $g(T(X)|\mu, \mu^2) = \frac{1}{\mu^n} e^{-\frac{\sum_{i=0}^{n} x_i^2}{2\mu^2}+\frac{\sum_{i=0}^{n} x_i}{\mu}-\frac{n}{2}}$

We have thus writen the joint pdf of the sample X as:

$$f(X_i, ..., X_n|\mu, \mu^2) = g(T(X)|\mu)h(X)$$

Therefore, by the **factorization theorem** $\{\sum_{i=0}^{n} x_i^2, \sum_{i=0}^{n} x_i\}$ is a sufficient statistic.

Next, we need to prove that this statistic is **not** complete. To do this we will find a linear combination of the statistic and prove that the following definition of a complete statistic does not hold:

$$E_\theta(g(T)) = 0, \forall\theta \rightarrow Pr(g(T) = 0; \theta) = 1\forall\theta$$

First, set $T_1 = \sum_{i=0}^{n} x_i$ and $T_2 = \sum_{i=0}^{n} x_i^2$. We will analyze the linear combination $2T_1^2 - (n+1)T_2$, first by finding its expected value:

$$E[2(\sum_{i=0}^{n} x_i)^2 - (n+1)\sum_{i=0}^{n} x_i^2] = 2E[(\sum_{i=0}^{n} x_i)^2] - (n+1)\sum_{i=0}^{n} E[x_i^2]$$

For the $E[x_i^2]$, since we known $E[X^2] = Var[X] + E[X]^2$, and since the variance of this function is μ^2, this can be rewritten as $(\mu^2 + \mu^2)$. For $E[\sum_{i=0}^{n} x_i)^2]$, we can use the same equation to rewrite this as $n\mu^2 + (n\mu)^2$. Thus, we can continue with rewriting the equation as follows:

$$\begin{aligned} 2E[(\sum_{i=0}^{n} x_i)^2] - (n+1)\sum_{i=0}^{n} E[x_i^2] &= 2(n\mu^2 + (n\mu)^2) - (n-1)n(\mu^2 + \mu^2) \\ &= (2n\mu^2 + 2n^2\mu^2) - (2n\mu^2 + 2n^2\mu^2) = 0 \end{aligned}$$

The expected value of this linear combination is 0. This implies that $Pr(g(T) = 0; \theta) = 1\forall\theta$ if the statistic is complete. So, if the statistic is complete, the following should be true:

$$Pr(2T_1^2 - (n+1)T_2 = 0) = 1$$

Therefore, we can rewrite the equality:

$$2T_1^2 = (n+1)T_2$$
$$T_2 = \frac{2T_1^2}{n+1}$$

The **Triangle Inequality** states that for any x, y $||x||^2 + ||y||^2 = ||x+y||^2$. Therefore, by the triangle inequality:

$$T_1^2 \leq T_2 = \frac{2T_1^2}{n+1}$$

However, when $n > 2$ this does not hold. Therefore, there is a contraction. **Thus, this linear combination is not trivially 0**. This proves that the statistic is sufficient, but not complete.

Exercise 9 (Regular exponential family)

Problem 1

We know the pdf of the Poisson distribution to be:

$$f(x|\lambda) = exp(-\lambda)\frac{\lambda^x}{x!}$$

This pdf can be rewritten as the following:

$$\begin{aligned} f(x|\lambda) &= exp(-\lambda)\lambda^x\frac{1}{x!} \\ &= \frac{1}{x!}exp(-\lambda)exp(log(\lambda^x)) \\ &= \frac{1}{x!}exp(-\lambda + xlog(\lambda)) \\ &= \frac{1}{x!}exp(xlog(\lambda) - \lambda) \end{aligned}$$

Let $\eta = log(\lambda)$, and let $B(\eta) = e^{log(\lambda)} = \lambda$, and let $T(X) = x$, and let $h(x) = \frac{1}{x!}$. Then the function can be rewritten in the following form:

$$\frac{1}{x!}exp(xlog(\lambda) - \lambda) = h(x)e^{\eta T(x) - B(\eta)}$$

Therefore, the Poisson distribution is a member of the regular exponential family.

Exercise 10 (Regular exponential family)

For a regular exponential family in the following form:

$$f(x|\eta) = h(x)exp(\eta T(x) - \beta(\eta))$$

the MGF of T(X) exists and from that the following properties are known:

$$E[T(X)|\eta] = \beta'(\eta)$$
$$Var[T(X)|\eta] = \beta''(\eta)$$

With these properties we can rearrange the formula for covariance:

$$\begin{aligned} Cov(T_i(X), T_j(X)) &= E[T_i(X)T_j(X)] - E[T_i]E[T_j] \\ &= Var[T_i(X)T_j(X)] + E[T_i]E[T_j] - E[T_i]E[T_j] \\ &= Var[T_i(X)T_j(X)] \\ &= \frac{\partial \beta(\eta)}{\partial \eta_i \partial \eta_j} \end{aligned}$$

Exercise 11 (Delta Method)

From CLT we know that $\bar{X}$ follows a normal distribution if n is large. Let $\hat{p} = \bar{X}$. Since we know the variance of a Bernoulli distribution is $p(1-p)$, the following holds:

$$\sqrt{n}[\hat{p} - p] \xrightarrow{D} N(0, p(1-p))$$

Let $g(p) = p(1-p)$ and thus $g'(p) = 1 - 2p$. Then by the Delta Method:

$$\begin{aligned} \sqrt{n}[g(\hat{p}) - g(p)] &\xrightarrow{d} N(0, p(1-p)g'(p)^2) \\ \sqrt{n}[(\hat{p}(1-\hat{p})) - (p(1-p))] &\xrightarrow{d} N(0, p(1-p)(1-2p)^2) \end{aligned}$$

Therefore, since $\hat{\tau} = \hat{p}(1\hat{p})$, we have proven that the approximate distribution of this estimator with a sufficiently large n is:

$$\hat{\tau} \sim N(p(1-p), \frac{p(1-p)(1-2p)^2}{n})$$

Information Theory

Exercise 12 (Joint Entropy)

Problem 1

We can compute the joint entropy using it's formula:

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} Pr(x,y) log_2(Pr(x,y)) \\ &= -\frac{1}{4}log_2(\frac{1}{4}) - \frac{1}{12}log_2(\frac{1}{12}) - \frac{1}{12}log_2(\frac{1}{12}) - \frac{1}{4}log_2(\frac{1}{4}) - \frac{1}{6}log_2(\frac{1}{6}) - \frac{1}{6}log_2(\frac{1}{6}) \\ &= -\frac{1}{2}log_2(\frac{1}{4}) - \frac{1}{6}log_2(\frac{1}{12}) - \frac{1}{3}log_2(\frac{1}{6}) \\ &= \frac{1}{2}log_2(4) + \frac{1}{6}log_2(12) + \frac{1}{3}log_2(6) \\ &= \frac{1}{2}log_2(2^2) + \frac{1}{6}log_2(3 * 2^2) + \frac{1}{3}log_2(3 * 2) \\ &= (\frac{1}{2} * 2 * 1) + (\frac{1}{6} * log_2(3) * 2 * 1) + (\frac{1}{3} * log_2(3) * 1) \\ &= \frac{5}{3} + \frac{log_2(3)}{2} \approx 2.459 \end{aligned}$$

Problem 2

First, to find the marginal distribution of X, add the probabilities for each X:

$$Pr(x=0) = \frac{1}{4} + \frac{1}{12} = \frac{1}{3}$$
$$Pr(x=1) = \frac{1}{12} + \frac{1}{4} = \frac{1}{3}$$
$$Pr(x=2) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

More formally, the marginal distribution of x is:

$$f_x = \begin{cases} \frac{1}{3}, & \text{if } x = 1 \\ \frac{1}{3}, & \text{if } x = 2 \\ \frac{1}{3}, & \text{if } x = 3 \end{cases}$$

To calculate the conditional entropy of $H(Y|X)$, we need to know the conditional probabilities of Y:

$$Pr(Y=y|X=x) = \frac{Pr(x,y)}{Pr(x)}$$
$$Pr(Y=0|X=0) = \frac{\frac{1}{4}}{\frac{1}{3}} = \frac{3}{4}$$
$$Pr(Y=0|X=1) = \frac{\frac{1}{12}}{\frac{1}{3}} = \frac{1}{4}$$
$$Pr(Y=0|X=2) = \frac{\frac{1}{6}}{\frac{1}{3}} = \frac{1}{2}$$
$$Pr(Y=1|X=0) = \frac{\frac{1}{12}}{\frac{1}{3}} = \frac{1}{4}$$
$$Pr(Y=1|X=1) = \frac{\frac{1}{4}}{\frac{1}{3}} = \frac{3}{4}$$
$$Pr(Y=1|X=2) = \frac{\frac{1}{6}}{\frac{1}{3}} = \frac{1}{2}$$

Next, the conditional entropy $H(Y|X)$ can be calculated using it's formula:

$$H(Y|X) = -\sum_{x\in X, y\in Y} Pr(x,y) log_2 Pr(Pr(Y=y|X=x))$$
$$= -\frac{log_2(\frac{3}{4})}{4} - \frac{log_2(\frac{1}{4})}{2} - \frac{log_2(\frac{1}{4})}{12} - \frac{log_2(\frac{3}{4})}{4} - \frac{log_2(\frac{1}{6})}{2} - \frac{log_2(\frac{1}{6})}{2}$$

This expression can be simplified similarly to problem one:

$$H(Y|X) = \frac{5}{3} - \frac{log_2(3)}{2} \approx 0.874$$

Problem 3

We can prove the two calculations above are correct using the chain rule:

$$\begin{aligned}H(Y|X) &= H(X,Y) - H(X)\\ \frac{5}{3} - \frac{log_2(3)}{2} &= \frac{5}{3} + \frac{log_2(3)}{2} - (-\sum_{x\in X} Pr(x)log_2(Pr(x)))\\ \frac{5}{3} - \frac{log_2(3)}{2} &= \frac{5}{3} + \frac{log_2(3)}{2} - 3(-\frac{1}{3}log_2(\frac{1}{3}))\\ \frac{5}{3} - \frac{log_2(3)}{2} &= \frac{5}{3} + \frac{log_2(3)}{2} - log_2(3)\\ \frac{5}{3} - \frac{log_2(3)}{2} &= \frac{5}{3} - \frac{log_2(3)}{2}\end{aligned}$$

The two sides of the equations are equal, thus the previous calculations must be correct.

Exercise 12 (Differential Entropy)

The pdf of a multivariate normal distribution is:

$$f(x;\mu,\mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k|\mathbf{\Sigma}|}}\exp(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu))$$

Based on the defintion of differential entropy we know:

$$H(X) = \int_{-\infty}^{+\infty} -log(f(x;\mu,\mathbf{\Sigma}))f(x;\mu,\mathbf{\Sigma})dx = -E[log(f(x;\mu,\mathbf{\Sigma}))]$$

We can expand out and solve this expected value:

$$\begin{aligned}H(X) &= E[log(\frac{1}{\sqrt{(2\pi)^k|\mathbf{\Sigma}|}}\exp(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)))]\\ &= -E[-\frac{k}{2}log(2\pi)] - E[-\frac{1}{2}log(|\mathbf{\Sigma}|)] - E[-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)]\\ &= \frac{k}{2}log(2\pi) + \frac{1}{2}log(|\mathbf{\Sigma}|) + \frac{1}{2}E[(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)]\end{aligned}$$

To find the remaining expected value, we'll use the *trace* properties discussed in the problem directions:

$$\begin{aligned}&= \frac{k}{2}log(2\pi) + \frac{1}{2}log(|\mathbf{\Sigma}|) + \frac{1}{2}E[trace((x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu))]\\ &= \frac{k}{2}log(2\pi) + \frac{1}{2}log(|\mathbf{\Sigma}|) + \frac{1}{2}trace(\mathbf{\Sigma}^{-1}E[(x-\mu)(x-\mu)^T])\\ &= \frac{k}{2}log(2\pi) + \frac{1}{2}log(|\mathbf{\Sigma}|) + \frac{1}{2}trace(\mathbf{\Sigma}^{-1}\mathbf{\Sigma})\\ &= \frac{k}{2}log(2\pi) + \frac{1}{2}log(|\mathbf{\Sigma}|) + \frac{1}{2}trace(I)\\ &= \frac{k}{2}log(2\pi) + \frac{1}{2}log(|\mathbf{\Sigma}|) + \frac{k}{2}\end{aligned}$$

Extra Credit

Application of Sufficency

Problem 1

Given we know the following equation is true for all random variables:

$$Var[X] = E[X^2] + E[X]^2$$

Then it follows that in a streaming setting, in order to get all the required information, you would need to save $\sum_{i=0}^{n} X_i$ and $\sum_{i=0}^{n} X_i^2$. Then with this information you would be able to reconstruct the variance and mean of the distribution. You would also need to save the value of n by implementing a counter alongside the summation of values.

In conclusion, you would need to save the following:

$$\{\sum_{i=0}^{n} x_i^2, \sum_{i=0}^{n} x_i, n\}$$

Huffman Coding

The algorithm for constructing a Huffman encoding is as follows:

1. Take two leaf nodes with the smallest probability/frequency.
2. Create a new internal node from those nodes whose weight is the some of the other nodes probabilities/frequencies.
3. Repeat on the new internal node with next leaf node that has equal probability/frequency to internal nodes weight.
4. I no equivalent leaf node exists, start over on the next leaf node.
5. Once complete, each branch is assigned a bit. It is customary is CS for the left bit to be 0 and the right to be 1.

6. The encoding for each word is found by traversing down the tree

With this algorithm, the following optimal tree can be constructed. (Note that optimal Huffman trees are not necessarily unique, so there are a few variations of this tree possible.)

When we traverse down the tree, the following encodings are found for each word. In this table I've also included the number of bits per encoding, and the frequency the word is used. Summing the product of word frequencies and their probability will get you the average number of bits needed to encode W. **The approximate number of bits needed is about 2.6.**

$$\text{Avg Bits} = \frac{52}{20} \approx 2.6$$

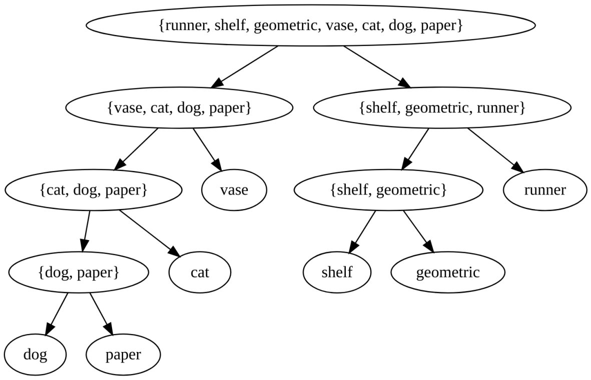


Figure 1: Optimal Huffman Tree. Assume every left branch indicates a 0 bit and every right branch indicates a 1.

Words	Frequency	Encoding	Bits	Bits.x.Frequency
runner	6	11	2	12
shelf	3	100	3	9
geometric	3	101	3	9
vase	4	1	2	8
cat	2	1	3	6
dog	1	0	4	4
paper	1	1	4	4