10-601 Machine Learning, Fall 2009: Homework 3

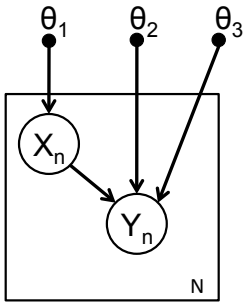
Due: Wednesday, October 7th, 10:30 am

Instructions There are 3 questions on this assignment worth the total of 140 points. Please hand in a hard copy at the beginning of the class. Refer to the webpage for policies regarding collaboration, due dates, and extensions.

1 Structured Density Estimation [40 Points]

In this problem you will have the opportunity to derive the maximum likelihood estimate (MLE) and maximum a posteriori (MAP) estimate for the parameters of a small structured model. You will also have the opportunity to derive the predictive distribution and become a better Bayesian.

Suppose you are given the simple Bayesian network in Fig. 1(a). You are told that the conditional probability tables (CPTs) for this model take the form of Table 1(b) and Table 1(c).



(a) Graphical Model

$\mathbf{P}(X=1)$	$\mathbf{P}(X=0)$
θ_1	$1-\theta_1$

(b) $\mathbf{P}(X)$

X	$\mathbf{P}(Y=1 \mid X)$	$\mathbf{P}(Y=0 \mid X)$
$X=1$	θ_2	$1-\theta_2$
$X=0$	θ_3	$1-\theta_3$

(c) $\mathbf{P}(Y \mid X)$

Figure 1: Graphical Model and Conditional Probability Tables

An expert on the study of X and Y (an XY-ologist) runs a large experiment and collects the data, $\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$. The marginal counts (e.g., a is the number of observations where $X = 1$ and $Y = 1$) for $\mathcal{D}$ are given by Table 1 where $a + b + c + d = N$. You will need to express your derivation using these counts.

	$Y=1$	$Y=0$
$X=1$	a	b
$X=0$	c	d

Table 1: Contingency Table

1.1 Deriving the MLE [15 Points]

We begin by deriving the MLE point estimates for $\{\theta_1, \theta_2, \theta_3\}$.

1. Show that the likelihood $\mathbf{P}\left(\mathcal{D} \mid \{\theta_1, \theta_2, \theta_3\}\right)$ can be expressed as:

$$\mathbf{P}\left(\mathcal{D} \mid \{\theta_1, \theta_2, \theta_3\}\right) = \theta_1^{a+b}(1-\theta_1)^{c+d}\theta_2^{a}(1-\theta_2)^{b}\theta_3^{c}(1-\theta_3)^{d}$$

★ **SOLUTION:** Since the data are iid, we can write the $\mathbf{P}\left(\mathcal{D} \mid \{\theta_1, \theta_2, \theta_3\}\right)$ as the product:

$$\begin{aligned}
\mathbf{P}\left(\mathcal{D} \mid \{\theta_1, \theta_2, \theta_3\}\right) &= \prod_{i=1}^{n} \mathbf{P}\left(X_i, Y_i \mid \{\theta_1, \theta_2, \theta_3\}\right) \\
&= \prod_{i=1}^{n} \mathbf{P}\left(Y_i \mid X_i \{\theta_1, \theta_2, \theta_3\}\right) \mathbf{P}\left(X_i \mid \{\theta_1, \theta_2, \theta_3\}\right) \\
&= \left(\prod_{i \in \{i \mid (x_i=1, y_i=1)\}} \cdots\right)\left(\prod_{i \in \{i \mid (x_i=1, y_i=0)\}} \cdots\right) \\
&\quad \left(\prod_{i \in \{i \mid (x_i=0, y_i=1)\}} \cdots\right)\left(\prod_{i \in \{i \mid (x_i=0, y_i=0)\}} \cdots\right) \\
&= (\theta_2\theta_1)^{a}\left((1-\theta_2)\theta_1\right)^{b}\left(\theta_3(1-\theta_1)\right)^{c}\left((1-\theta_3)(1-\theta_1)\right)^{d} \\
&= \theta_1^{a+b}(1-\theta_1)^{c+d}\theta_2^{a}(1-\theta_2)^{b}\theta_3^{c}(1-\theta_3)^{d}
\end{aligned}$$

2. Take the log of the likelihood to construct the log-likelihood, $\mathcal{L}\left(\theta_1, \theta_2, \theta_3\right) = \log\left(\mathbf{P}\left(\mathcal{D} \mid \{\theta_1, \theta_2, \theta_3\}\right)\right)$, as a sum of terms.

★ **SOLUTION:** This is simply achieved by taking the log of the above likelihood. Notice that the product decomposes into a more manageable sum:

$$\begin{aligned}
\mathcal{L}\left(\theta_1, \theta_2, \theta_3\right) &= \log\left(\mathbf{P}\left(\mathcal{D} \mid \{\theta_1, \theta_2, \theta_3\}\right)\right) \\
&= \log\left(\theta_1^{a+b}(1-\theta_1)^{c+d}\theta_2^{a}(1-\theta_2)^{b}\theta_3^{c}(1-\theta_3)^{d}\right) \\
&= (a+b)\log(\theta_1) + (c+d)\log(1-\theta_1) + a\log(\theta_2) + \\
&\quad b\log(1-\theta_2) + c\log(\theta_3) + d\log(1-\theta_3)
\end{aligned}$$

3. Why is the following expression relevant when trying to find parameters that maximize the likelihood?

$$\arg\max_x \log\left(f(x)\right) = \arg\max_x f(x)$$

★ **SOLUTION:** Because the $\log()$ function is a monotonically increasing over the positive domain we can maximize $\log(f(x))$ where $f(x) \geq 0$ and the point that maximizes $f(x)$ will also maximize $\log(f(x))$. This allows us to reduce the analytically more challenging maximization over a product to the simpler maximization over a sum.

4. Write the partial derivatives of the log-likelihood with respect to each parameter (i.e., $\frac{\partial \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_1}$, $\frac{\partial \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_2}$, and $\frac{\partial \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_3}$).

★ **SOLUTION:** Here we simply compute each of the partial derivatives:

$$\begin{aligned}
\frac{\partial \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_1} &= \frac{a+b}{\theta_1} - \frac{c+d}{1-\theta_1} \\
\frac{\partial \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_2} &= \frac{a}{\theta_2} - \frac{b}{1-\theta_2} \\
\frac{\partial \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_3} &= \frac{c}{\theta_3} - \frac{d}{1-\theta_3}
\end{aligned}$$

5. Do the partial derivatives with respect to each parameter depend on the other parameters? What does this say with respect to computing the maximizing joint assignment for $\{\theta_1, \theta_2, \theta_3\}$?

★ **SOLUTION:** The partial derivatives do not share parameter values and therefore the likelihood can be optimized independently with respect to each parameter in $\{\theta_1, \theta_2, \theta_3\}$

6. Write the second partial derivatives ($\frac{\partial^2 \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_1^2}$, $\frac{\partial^2 \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_2^2}$, and $\frac{\partial^2 \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_3^2}$) of the log-likelihood.

★ **SOLUTION:** Here we simply compute each of the second partial derivatives:

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_1^2} &= -\frac{a+b}{\theta_1^2} - \frac{c+d}{(1-\theta_1)^2} \\
\frac{\partial^2 \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_2^2} &= -\frac{a}{\theta_2^2} - \frac{b}{(1-\theta_2)^2} \\
\frac{\partial^2 \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_3^2} &= -\frac{c}{\theta_3^2} - \frac{d}{(1-\theta_3)^2}
\end{aligned}$$

7. Using your answer to the previous question, argue that this function is concave with respect to each parameter in $\{\theta_1, \theta_2, \theta_3\}$ individually. How is concavity relevant to the problem of maximization?

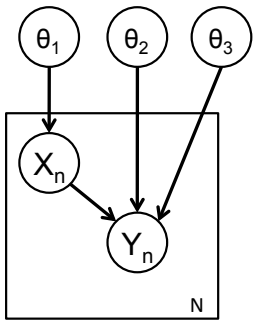
★ **SOLUTION:** The second derivatives of $\mathcal{L}(\theta_1, \theta_2, \theta_3)$ are always negative (for nonzero counts) and therefore the function is concave with respect to each of the parameters individually. Therefore, there is a single unique maximizing assignment.

8. Set each of the first partial derivatives (i.e., $\frac{\partial \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_1}$, $\frac{\partial \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_2}$, and $\frac{\partial \mathcal{L}(\theta_1, \theta_2, \theta_3)}{\partial \theta_3}$) equal to zero and solve for the maximizing assignment. This your maximum likelihood estimate. Comment on the form of the estimate in one sentence.

★ **SOLUTION:** With basic algebra we can solve for each of the parameters to obtain:

$$\begin{aligned}
\theta_1 &= \frac{a+b}{a+b+c+d} \\
\theta_2 &= \frac{a}{a+b} \\
\theta_3 &= \frac{c}{c+d}
\end{aligned}$$

These are exactly the empirical counts used in the direct counting method.



(a) Graphical Model

$$\begin{aligned}\theta_1 &\sim \text{Beta}(\alpha_1, \beta_1)\\ \theta_2 &\sim \text{Beta}(\alpha_2, \beta_2)\\ \theta_3 &\sim \text{Beta}(\alpha_3, \beta_3)\end{aligned}$$

(b) Priors

Figure 2: Graphical model treating $\{\theta_1, \theta_2, \theta_3\}$ as random variables with priors.

1.2 Deriving the Maximum A Posteriori (MAP) Estimate [10 Points]

If you were a pragmatic Bayesian you would put a prior on the parameters $\{\theta_1, \theta_2, \theta_3\}$ and compute the MAP estimate instead of the MLE. You talk to the expert on X and Y and she gives you the Beta priors in Fig. 2 where α_1, β_1, α_2, β_2, α_3, and β_3 are all positive integers. Recall from class that the $\text{Beta}(\theta \mid \alpha, \beta)$ takes the form:

$$\text{Beta}(\theta \mid \alpha, \beta) = \frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

1. Write the posterior $\mathbf{P}(\{\theta_1, \theta_2, \theta_3\} \mid \mathcal{D})$. You may want to use the likelihood from the previous question.

   ★ **SOLUTION:** Using Bayes Rule and substituting the form for the Beta prior:

$$\begin{aligned}\mathbf{P}(\theta_1, \theta_2, \theta_3 \mid \mathcal{D}) &\propto \mathbf{P}(\mathcal{D} \mid \theta_1, \theta_2, \theta_3)\,\mathbf{P}(\theta_1, \theta_2, \theta_3)\\ &\propto \mathbf{P}(\mathcal{D} \mid \theta_1, \theta_2, \theta_3)\,\mathbf{P}(\theta_1)\,\mathbf{P}(\theta_2)\,\mathbf{P}(\theta_3)\\ &\propto \theta_1^{a+b}(1-\theta_1)^{c+d}\theta_2^{a}(1-\theta_2)^{b}\theta_3^{c}(1-\theta_3)^{d}\\ &\quad \left(\theta_1^{\alpha_1-1}(1-\theta_1)^{\beta_1-1}\right)\left(\theta_2^{\alpha_2-1}(1-\theta_2)^{\beta_2-1}\right)\left(\theta_3^{\alpha_3-1}(1-\theta_3)^{\beta_3-1}\right)\\ &\propto \theta_1^{a+b+\alpha_1-1}(1-\theta_1)^{c+d+\beta_1-1}\theta_2^{a+\alpha_2-1}(1-\theta_2)^{b+\beta_2-1}\theta_3^{c+\alpha_3-1}(1-\theta_3)^{d+\beta_3-1}\end{aligned}$$

   We will not need the log-partition function (the normalizing constant) since it only depends on the hyper-parameters. However we could quickly derive it by using the product of Beta normalizing constants.

2. Take the log of the posterior to construct the log-posterior as a sum of terms.

   ★ **SOLUTION:** Doing the algebra:

$$\begin{aligned}\log(\mathbf{P}(\theta_1, \theta_2, \theta_3 \mid \mathcal{D})) &= (a+b+\alpha_1-1)\log(\theta_1) + (c+d+\beta_1-1)\log(1-\theta_1) +\\ &\quad (a+\alpha_2-1)\log(\theta_2) + (b+\beta_2-1)\log(1-\theta_2) +\\ &\quad (c+\alpha_3-1)\log(\theta_3) + (d+\beta_3-1)\log(1-\theta_3) -\\ &\quad \log(Z(\alpha, \beta))\end{aligned}$$

   Notice that we have kept around the $\log(Z(\alpha, \beta))$ normalizing constant. This is not necessary but it enables us to use strict equality.

3. Write the partial derivatives of the log-posterior with respect to θ_1, θ_2, and θ_3.

$$\begin{aligned}
\frac{\partial \log\left(\mathbf{P}\left(\{\theta_1, \theta_2, \theta_3\} \mid \mathcal{D}\right)\right)}{\partial \theta_1} &= \\
\frac{\partial \log\left(\mathbf{P}\left(\{\theta_1, \theta_2, \theta_3\} \mid \mathcal{D}\right)\right)}{\partial \theta_2} &= \\
\frac{\partial \log\left(\mathbf{P}\left(\{\theta_1, \theta_2, \theta_3\} \mid \mathcal{D}\right)\right)}{\partial \theta_3} &=
\end{aligned}$$

★ **SOLUTION:** Similar to the derivatives of the likelihood we get:

$$\begin{aligned}
\frac{\partial \log\left(\mathbf{P}\left(\{\theta_1, \theta_2, \theta_3\} \mid \mathcal{D}\right)\right)}{\partial \theta_1} &= \frac{a+b+\alpha_1-1}{\theta_1} - \frac{c+d+\beta_1-1}{1-\theta_1} \\
\frac{\partial \log\left(\mathbf{P}\left(\{\theta_1, \theta_2, \theta_3\} \mid \mathcal{D}\right)\right)}{\partial \theta_2} &= \frac{a+\alpha_2-1}{\theta_2} - \frac{b+\beta_2-1}{1-\theta_2} \\
\frac{\partial \log\left(\mathbf{P}\left(\{\theta_1, \theta_2, \theta_3\} \mid \mathcal{D}\right)\right)}{\partial \theta_3} &= \frac{c+\alpha_3-1}{\theta_3} - \frac{d+\beta_3-1}{1-\theta_3}
\end{aligned}$$

4. Is $\log\left(\mathbf{P}\left(\{\theta_1, \theta_2, \theta_3\} \mid \mathcal{D}\right)\right)$ concave with respect to θ_1, θ_2, and θ_3 individually?

★ **SOLUTION:** This is log-concave for $\alpha, \beta \geq 1$. This can be seen again by taking the second derivatives. However, by analogy to the MLE we can quickly see that the second derivatives are always negative.

5. Do the partial derivatives with respect to each parameter depend on the other parameters? What does this say with respect to computing the maximizing joint assignment for $\{\theta_1, \theta_2, \theta_3\}$?

★ **SOLUTION:** No. The partial derivatives are independent with respect to the parameters, which as before enables us to independently solve for the optimal value for each parameter.

6. Set each of the partial derivatives equal to zero and solve for the maximizing assignment to each parameter $\{\theta_1, \theta_2, \theta_3\}$. This is your MAP estimate. Comment on the form of the estimate.

★ **SOLUTION:** Again with some basic algebra we find:

$$\begin{aligned}
\theta_1 &= \frac{a+b+\alpha_1-1}{a+b+c+d+\alpha_1+\beta_1-2} \\
\theta_2 &= \frac{a+\alpha_2-1}{a+b+\alpha_2+\beta_2-2} \\
\theta_3 &= \frac{c+\alpha_3-1}{c+d+\alpha_3+\beta_3-2}
\end{aligned}$$

This is analogous to the MLE estimates where we have added the additional "counts" given by the α and β parameters. Notice to obtain Laplace smoothing we would need to set $\alpha = \beta = 2$.

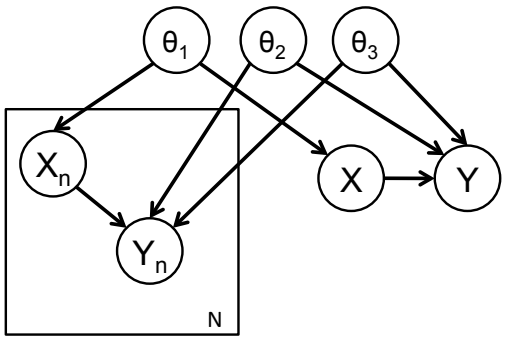


Figure 3: The model for the predictive distribution on X and Y.

1.3 Deriving the Predictive Distribution [15 Points]

A proper Bayesian would not likely settle for a MAP estimate. Instead he, she, or possibly some distribution over genders would probably marginalize away the unknown parameters to construct the posterior predictive distribution. Here we will derive the posterior predictive distributions $\mathbf{P}(X \mid \mathcal{D})$, $\mathbf{P}(Y \mid X = 0, \mathcal{D})$, and $\mathbf{P}(Y \mid X = 1, \mathcal{D})$. We will use the priors on $\{\theta_1, \theta_2, \theta_3\}$ obtained from the XY-ologist in the previous question.

1.3.1 Posterior Predictive Distribution for X

1. Using the equation for the posterior $\mathbf{P}(\{\theta_1, \theta_2, \theta_3\} \mid \mathcal{D})$ show that:

$$\mathbf{P}(\{\theta_1, \theta_2, \theta_3\} \mid \mathcal{D}) = \mathbf{P}(\theta_1 \mid \mathcal{D})\,\mathbf{P}(\theta_2 \mid \mathcal{D})\,\mathbf{P}(\theta_3 \mid \mathcal{D})$$

The previous means that θ_1, θ_2, and θ_3 are conditionally independent given the data. Can you show any of these conditional independencies using only d-separation? (Consider showing (i) $\theta_1 \perp \theta_2 \mid \mathcal{D}$; (ii) $\theta_1 \perp \theta_3 \mid \mathcal{D}$; and (iii) $\theta_2 \perp \theta_3 \mid \mathcal{D}$.)

★ **SOLUTION:** From the MAP derivation we have:

$$\begin{aligned}
\mathbf{P}(\theta_1, \theta_2, \theta_3 \mid \mathcal{D}) &\propto \left[\theta_1^{a+b+\alpha_1-1}(1-\theta_1)^{c+d+\beta_1-1}\right]\left[\theta_2^{a+\alpha_2-1}(1-\theta_2)^{b+\beta_2-1}\right]\left[\theta_3^{c+\alpha_3-1}(1-\theta_3)^{d+\beta_3-1}\right] \\
\mathbf{P}(\theta_1 \mid \mathcal{D}) &\propto \theta_1^{a+b+\alpha_1-1}(1-\theta_1)^{c+d+\beta_1-1} \\
\mathbf{P}(\theta_2 \mid \mathcal{D}) &\propto \theta_2^{a+\alpha_2-1}(1-\theta_2)^{b+\beta_2-1} \\
\mathbf{P}(\theta_3 \mid \mathcal{D}) &\propto \theta_3^{c+\alpha_3-1}(1-\theta_3)^{d+\beta_3-1}
\end{aligned}$$

Using d-separation we can show that $\theta_1 \perp \theta_2 \mid \mathcal{D}$ and $\theta_1 \perp \theta_3 \mid \mathcal{D}$. However we cannot show $\theta_2 \perp \theta_3 \mid \mathcal{D}$ from the graph structure alone. This is conditional independence is derived from the conditional probability tables and is referred to as context specific independence.

2. Using d-separation show that:

$$\mathbf{P}(X, \theta_1 \mid \mathcal{D}) = \mathbf{P}(X \mid \theta_1)\,\mathbf{P}(\theta_1 \mid \mathcal{D})$$

★ **SOLUTION:** By the definition of the conditional we can write:

$$\mathbf{P}(X, \theta_1 \mid \mathcal{D}) = \mathbf{P}(X \mid \theta_1, \mathcal{D})\,\mathbf{P}(\theta_1 \mid \mathcal{D})$$

Then we notice that the new X is conditionally independent of the $\mathcal{D}$ given θ_1 and so we can write:

$$\mathbf{P}(X \mid \theta_1, \mathcal{D}) = \mathbf{P}(X \mid \theta_1)$$

3. Argue that for some normalizing constant Z:

$$\begin{aligned}
\mathbf{P}\left(X=1\,|\,\mathcal{D}\right) &= \frac{1}{Z}\int_0^1 \theta_1 \left(\theta_1^{a+b}(1-\theta_1)^{c+d}\right)\left(\theta_1^{\alpha_1-1}(1-\theta_1)^{\beta_1-1}\right)d\theta_1 \\
\mathbf{P}\left(X=0\,|\,\mathcal{D}\right) &= \frac{1}{Z}\int_0^1 (1-\theta_1) \left(\theta_1^{a+b}(1-\theta_1)^{c+d}\right)\left(\theta_1^{\alpha_1-1}(1-\theta_1)^{\beta_1-1}\right)d\theta_1
\end{aligned}$$

★ **SOLUTION:** We begin by marginalizing the joint $\mathbf{P}\left(X,\theta_1\,|\,\mathcal{D}\right)$

$$\begin{aligned}
\mathbf{P}\left(X=1\,|\,\mathcal{D}\right) &= \int_0^1 \mathbf{P}\left(X=1,\theta_1\,|\,\mathcal{D}\right)d\theta_1 \\
&= \int_0^1 \mathbf{P}\left(X=1\,|\,\theta_1\right)\mathbf{P}\left(\theta_1\,|\,\mathcal{D}\right)d\theta_1 \\
&= \frac{1}{Z}\int_0^1 [\theta_1]\left[\theta_1^{a+b+\alpha_1-1}(1-\theta_1)^{c+d+\beta_1-1}\right]d\theta_1 \\
\mathbf{P}\left(X=0\,|\,\mathcal{D}\right) &= \int_0^1 \mathbf{P}\left(X=0,\theta_1\,|\,\mathcal{D}\right)d\theta_1 \\
&= \int_0^1 \mathbf{P}\left(X=0\,|\,\theta_1\right)\mathbf{P}\left(\theta_1\,|\,\mathcal{D}\right)d\theta_1 \\
&= \frac{1}{Z}\int_0^1 [(1-\theta_1)]\left[\theta_1^{a+b+\alpha_1-1}(1-\theta_1)^{c+d+\beta_1-1}\right]d\theta_1
\end{aligned}$$

We can simplify this expression further which will be helpful in the next part:

$$\begin{aligned}
\mathbf{P}\left(X=1\,|\,\mathcal{D}\right) &= \frac{1}{Z}\int_0^1 \theta_1^{a+b+\alpha_1}(1-\theta_1)^{c+d+\beta_1-1}d\theta_1 \\
\mathbf{P}\left(X=0\,|\,\mathcal{D}\right) &= \frac{1}{Z}\int_0^1 \theta_1^{a+b+\alpha_1-1}(1-\theta_1)^{c+d+\beta_1}d\theta_1
\end{aligned}$$

4. Using the following identities where we assume r and q are positive integers:

$$\begin{aligned}
\int_0^1 z^{q-1}(1-z)^{r-1}dz &= \frac{(q-1)!(r-1)!}{(q+r-1)!} \\
r! &= r(r-1)!
\end{aligned}$$

derive the following expressions

$$\begin{aligned}
\mathbf{P}\left(X=1\,|\,\mathcal{D}\right) &= \frac{a+b+\alpha_1}{N+\alpha_1+\beta_1} \\
\mathbf{P}\left(X=0\,|\,\mathcal{D}\right) &= \frac{c+d+\beta_1}{N+\alpha_1+\beta_1}
\end{aligned}$$

This is your posterior predictive distribution for X. Comment on how it differs from the MAP estimate for θ_1 Which one gives you a 'smoother' prediction?

★ **SOLUTION:** The trick to this question is simply taking the above integrals:

$$
\begin{aligned}
\mathbf{P}\left(X=1 \mid \mathcal{D}\right) &= \left(\frac{1}{Z}\right) \frac{((a+b+\alpha_1+1)-1)!((c+d+\beta_1)-1)!}{((a+b+\alpha_1+1)+(c+d+\beta_1)-1)!} \\
&= \left(\frac{1}{Z}\right) \frac{(a+b+\alpha_1)!(c+d+\beta_1-1)!}{(N+\alpha_1+\beta_1)!} \\
\mathbf{P}\left(X=0 \mid \mathcal{D}\right) &= \left(\frac{1}{Z}\right) \frac{((a+b+\alpha_1)-1)!((c+d+\beta_1+1)-1)!}{((a+b+\alpha_1)+(c+d+\beta_1+1)-1)!} \\
&= \left(\frac{1}{Z}\right) \frac{(a+b+\alpha_1-1)!(c+d+\beta_1)!}{(N+\alpha_1+\beta_1)!}
\end{aligned}
$$

Using the identity $r(r-1)!$ we can further reduce terms:

$$
\begin{aligned}
\mathbf{P}\left(X=1 \mid \mathcal{D}\right) &= \left(\frac{1}{Z}\frac{(a+b+\alpha_1-1)!(c+d+\beta_1-1)!}{(N+\alpha_1+\beta_1-1)!}\right)\frac{a+b+\alpha_1}{N+\alpha_1+\beta_1} \\
\mathbf{P}\left(X=0 \mid \mathcal{D}\right) &= \left(\frac{1}{Z}\frac{(a+b+\alpha_1-1)!(c+d+\beta_1-1)!}{(N+\alpha_1+\beta_1-1)!}\right)\frac{c+d+\beta_1}{N+\alpha_1+\beta_1}
\end{aligned}
$$

We can see that setting Z:

$$
\begin{aligned}
Z &= \mathbf{P}\left(X=1 \mid \mathcal{D}\right) + \mathbf{P}\left(X=0 \mid \mathcal{D}\right) \\
&= \frac{(a+b+\alpha_1-1)!(c+d+\beta_1-1)!}{(N+\alpha_1+\beta_1-1)!}
\end{aligned}
$$

we ensure proper normalization. The posterior predictive distribution provides 1 additional count over MAP estimation and is therefore slightly smoother. This is exactly Laplace smoothing.

1.3.2 Posterior Predictive Distribution for Y

1. Using d-separation argue that:

$$\mathbf{P}\left(Y, \theta_2, \theta_3 \mid X, \mathcal{D}\right) = \mathbf{P}\left(Y \mid X, \theta_2, \theta_3\right)\mathbf{P}\left(\theta_2, \theta_3 \mid \mathcal{D}\right)$$

★ **SOLUTION:** Using the definition of the conditional probability we get:

$$\mathbf{P}\left(Y, \theta_2, \theta_3 \mid X, \mathcal{D}\right) = \mathbf{P}\left(Y \mid \theta_2, \theta_3, X, \mathcal{D}\right)\mathbf{P}\left(\theta_2, \theta_3 \mid X, \mathcal{D}\right)$$

By d-separation Y is conditionally independent of the data given θ_2, θ_3, and X.

$$\mathbf{P}\left(Y \mid \theta_2, \theta_3, X, \mathcal{D}\right) = \mathbf{P}\left(Y \mid \theta_2, \theta_3, X\right)$$

By d-separation θ_2 and θ_3 are conditionally independent of X given the data.

$$\mathbf{P}\left(\theta_2, \theta_3 \mid X, \mathcal{D}\right) = \mathbf{P}\left(\theta_2, \theta_3 \mid \mathcal{D}\right)$$

2. Using the results from the previous questions argue that:

$$\mathbf{P}\left(Y \mid X, \mathcal{D}\right) = \int_0^1 \int_0^1 \mathbf{P}\left(Y \mid X, \theta_2, \theta_3\right)\mathbf{P}\left(\theta_2 \mid \mathcal{D}\right)\mathbf{P}\left(\theta_3 \mid \mathcal{D}\right) d\theta_2 d\theta_3$$

★ **SOLUTION:** We begin by writing $\mathbf{P}(Y \mid X, \mathcal{D})$ as a marginal of the joint $\mathbf{P}(Y, \theta_2, \theta_3 \mid X, \mathcal{D})$

$$\mathbf{P}(Y \mid X, \mathcal{D}) = \int_0^1 \int_0^1 \mathbf{P}(Y\theta_x, \theta_3 \mid X, \mathcal{D})\, d\theta_2 d\theta_3$$

Then using the answer to the previous questions we get:

$$\begin{aligned}
\mathbf{P}(Y \mid X, \mathcal{D}) &= \int_0^1 \int_0^1 \mathbf{P}(Y\theta_x, \theta_3 \mid X, \mathcal{D})\, d\theta_2 d\theta_3 \\
&= \int_0^1 \int_0^1 \mathbf{P}(Y \mid X, \theta_2, \theta_3)\, \mathbf{P}(\theta_2, \theta_3 \mid \mathcal{D})\, d\theta_2 d\theta_3 \\
&= \int_0^1 \int_0^1 \mathbf{P}(Y \mid X, \theta_2, \theta_3)\, \mathbf{P}(\theta_2 \mid \mathcal{D})\, \mathbf{P}(\theta_3 \mid \mathcal{D})\, d\theta_2 d\theta_3
\end{aligned}$$

The last step comes from the derivation of the factorized posterior in 1.3.1 where we used context specific independence.

3. We can eliminate the double integral by evaluating it for each assignment to X separately. Using the definition of $\mathbf{P}(Y \mid X, \theta_2, \theta_3)$ show that:

$$\begin{aligned}
\mathbf{P}(Y \mid X = 1, \mathcal{D}) &= \int_0^1 \mathbf{P}(Y \mid X = 1, \theta_2)\, \mathbf{P}(\theta_2 \mid \mathcal{D})\, d\theta_2 \\
\mathbf{P}(Y \mid X = 0, \mathcal{D}) &= \int_0^1 \mathbf{P}(Y \mid X = 0, \theta_3)\, \mathbf{P}(\theta_3 \mid \mathcal{D})\, d\theta_3
\end{aligned}$$

The type of simplification we used above is called *context specific independence*, because it is facilitated through specific settings of the parents (these settings are called "contexts").

★ **SOLUTION:** Notice that when we restrict X to take a particular value:

$$\begin{aligned}
\mathbf{P}(Y \mid X = 1, \theta_2, \theta_3) &= \theta_2^y (1 - \theta_2)^{1-y} \\
&= \mathbf{P}(Y \mid X = 1, \theta_2) \\
\mathbf{P}(Y \mid X = 0, \theta_2, \theta_3) &= \theta_3^y (1 - \theta_3)^{1-y} \\
&= \mathbf{P}(Y \mid X = 0, \theta_3)
\end{aligned}$$

the conditional probability $\mathbf{P}(Y \mid X, \theta_2, \theta_3)$ further factors. Using this we obtain:

$$\begin{aligned}
\mathbf{P}(Y \mid X = 1, \mathcal{D}) &= \int_0^1 \int_0^1 \mathbf{P}(Y \mid X = 1, \theta_2, \theta_3)\, \mathbf{P}(\theta_2 \mid \mathcal{D})\, \mathbf{P}(\theta_3 \mid \mathcal{D})\, d\theta_2 d\theta_3 \\
&= \int_0^1 \mathbf{P}(Y \mid X = 1, \theta_2)\, \mathbf{P}(\theta_2 \mid \mathcal{D}) \int_0^1 \mathbf{P}(\theta_3 \mid \mathcal{D})\, d\theta_3 d\theta_2 \\
&= \int_0^1 \mathbf{P}(Y \mid X = 1, \theta_2)\, \mathbf{P}(\theta_2 \mid \mathcal{D})\, d\theta_2 \\
\mathbf{P}(Y \mid X = 0, \mathcal{D}) &= \int_0^1 \int_0^1 \mathbf{P}(Y \mid X = 0, \theta_2, \theta_3)\, \mathbf{P}(\theta_2 \mid \mathcal{D})\, \mathbf{P}(\theta_3 \mid \mathcal{D})\, d\theta_2 d\theta_3 \\
&= \int_0^1 \mathbf{P}(Y \mid X = 0, \theta_3)\, \mathbf{P}(\theta_3 \mid \mathcal{D}) \int_0^1 \mathbf{P}(\theta_2 \mid \mathcal{D})\, d\theta_2 d\theta_3 \\
&= \int_0^1 \mathbf{P}(Y \mid X = 0, \theta_3)\, \mathbf{P}(\theta_3 \mid \mathcal{D})\, d\theta_3
\end{aligned}$$

4. Evaluate the integrals to show that:

$$\begin{aligned}\mathbf{P}\left(Y=1\,|\,X=1,\mathcal{D}\right) &= \frac{a+\alpha_2}{a+b+\alpha_2+\beta_2}\\ \mathbf{P}\left(Y=1\,|\,X=0,\mathcal{D}\right) &= \frac{c+\alpha_3}{c+d+\alpha_3+\beta_3}\end{aligned}$$

This is your predictive distribution for $Y\,|\,X$. Explain how this relates to Laplace smoothing.

★ **SOLUTION:** We will first focus on the case for $X=1$

$$\begin{aligned}\mathbf{P}\left(Y=1\,|\,X=1,\mathcal{D}\right) &= \frac{1}{Z}\int_0^1 \mathbf{P}\left(Y=1\,|\,X=1,\theta_2\right)\mathbf{P}\left(\theta_2\,|\,\mathcal{D}\right)d\theta_2\\ &= \frac{1}{Z}\int_0^1 \theta_2\theta_2^{a+\alpha_2-1}(1-\theta_2)^{b+\beta_2-1}d\theta_2\\ &= \frac{1}{Z}\int_0^1 \theta_2^{a+\alpha_2}(1-\theta_2)^{b+\beta_2-1}d\theta_2\\ &= \frac{1}{Z}\frac{((a+\alpha_2+1)-1)!((b+\beta_2)-1)!}{((a+\alpha_2+1)+(b+\beta_2)-1)!}\\ &= \frac{1}{Z}\frac{(a+\alpha_2)!(b+\beta_2-1)!}{(a+\alpha_2+b+\beta_2)!}\\ \mathbf{P}\left(Y=0\,|\,X=1,\mathcal{D}\right) &= \frac{1}{Z}\int_0^1 \mathbf{P}\left(Y=0\,|\,X=1,\theta_2\right)\mathbf{P}\left(\theta_2\,|\,\mathcal{D}\right)d\theta_2\\ &= \frac{1}{Z}\int_0^1 (1-\theta_2)\theta_2^{a+\alpha_2-1}(1-\theta_2)^{b+\beta_2-1}d\theta_2\\ &= \frac{1}{Z}\int_0^1 \theta_2^{a+\alpha_2-1}(1-\theta_2)^{b+\beta_2}d\theta_2\\ &= \frac{1}{Z}\frac{((a+\alpha_2)-1)!((b+\beta_2+1)-1)!}{((a+\alpha_2+1)+(b+\beta_2)-1)!}\\ &= \frac{1}{Z}\frac{(a+\alpha_2-1)!(b+\beta_2)!}{(a+\alpha_2+b+\beta_2)!}\end{aligned}$$

We then obtain the normalizing constant:

$$Z = \frac{(a+\alpha_2-1)!(b+\beta_2-1)!}{(a+\alpha_2+b+\beta_2-1)!}$$

Which when we divide by Z we get the desired result:

$$\mathbf{P}\left(Y=1\,|\,X=1,\mathcal{D}\right)=\frac{a+\alpha_2}{a+b+\alpha_2+\beta_2}$$

Repeating the same procedure for $X = 0$ (by analogy):

$$\begin{aligned}
\mathbf{P}(Y = 1 \mid X = 0, \mathcal{D}) &= \frac{1}{Z}\int_0^1 \mathbf{P}(Y = 1 \mid X = 0, \theta_3)\,\mathbf{P}(\theta_3 \mid \mathcal{D})\,d\theta_3 \\
&= \frac{1}{Z}\int_0^1 \theta_3\theta_3^{c+\alpha_3-1}(1-\theta_3)^{d+\beta_3-1}d\theta_3 \\
&= \frac{1}{Z}\frac{(c+\alpha_3)!(d+\beta_3-1)!}{(c+\alpha_3+d+\beta_3)!} \\
\mathbf{P}(Y = 0 \mid X = 0, \mathcal{D}) &= \frac{1}{Z}\int_0^1 \mathbf{P}(Y = 0 \mid X = 0, \theta_3)\,\mathbf{P}(\theta_3 \mid \mathcal{D})\,d\theta_3 \\
&= \frac{1}{Z}\int_0^1 (1-\theta_3)\theta_3^{c+\alpha_3-1}(1-\theta_3)^{d+\beta_3-1}d\theta_3 \\
&= \frac{1}{Z}\frac{(c+\alpha_3-1)!(d+\beta_3)!}{(c+\alpha_3+d+\beta_3)!}
\end{aligned}$$

We then obtain the normalizing constant:

$$Z = \frac{(c+\alpha_3-1)!(d+\beta_3-1)!}{(c+\alpha_3+d+\beta_3-1)!}$$

Which when we divide by Z we get the desired result:

$$\mathbf{P}(Y = 1 \mid X = 0, \mathcal{D}) = \frac{c+\alpha_3}{c+d+\alpha_3+\beta_3}$$

This is again the Laplace smoothing with $\alpha = \beta = 1$ which is a uniform prior.

5. Which parameter estimation method would likely be most robust when there is limited data: MLE, MAP, or Posterior Predictive Distribution?

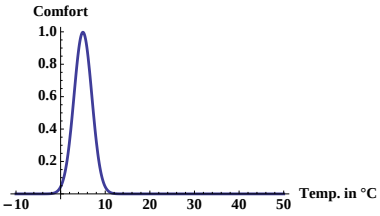
★ **SOLUTION:** The posterior predictive distribution introduces the most smoothing and therefore reduces variance at the cost of increased bias. In situations where data is sparse the variance is often large, and the posterior predictive distribution would be the preferred estimator.

2 Suitcase Packing [15 Points]

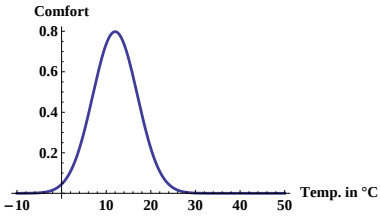
As any computer scientist will tell you, packing a suitcase is really difficult. Even if we ignore the substantial challenge of fitting everything, we are still left with the more mundane challenge of selecting the right clothing for the occasion. Naturally, we will ignore less important criteria like color and style, and focus on important characteristics like comfort with respect to temperature.

If the destination is very warm then we might want to pack clothing that will help us stay cool like swim trunks and sun-glasses. Conversely, if the destination is cold, then we would definitely want to pack gloves, a scarf, and some skis. If the destination is mild then we may want to bring pants and a light coat. What do we pack when the destination is Pittsburgh and packing for the average temperature won't suffice? To help make your life easier, we have prepared 4 prepacked suitcases at no extra charge. Naturally, we have included your utility curves for each suitcase in Fig. 4. The utility curves describe how comfortable you will be at each temperature if you select that suitcase.

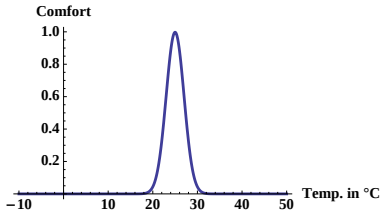
In this question we will investigate the disadvantages of point estimates and try to illustrate the importance of working with the posterior distribution. For this question you will not need to do any calculations or derive equations; instead, we ask you to provide concise explanations for your decisions.



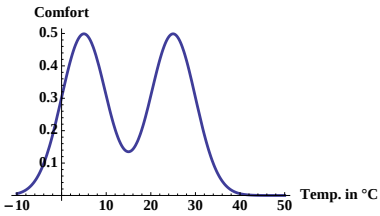
(a) Suitcase for Cold Weather



(b) Suitcase for Mild Weather



(c) Suitcase for Hot Weather



(d) Suitcase for Mixed Conditions

Figure 4: The comfort zones for each suitcase. Notice that the suitcase for mixed conditions is not quite as comfortable when its hot or cold as the respective suitcases for hot and cold weather.

2.1 The Average Temperature

As part of your preliminary planning phase you check the weather at your destination and discover that the average temperature is $15°C$. Which suitcase do you choose and why?

★ **SOLUTION:** *(b).* Because you are told that the average weather is mild and you do not have any additional information you would likely choose the mild weather suitcase. This suitcase will make you most comfortable if the weather is around $15°C$.

2.2 The Most Common Temperature

You call a friend who frequently visits your destination and he says that it is usually $25°C$. Which suitcase do you choose and why?

★ **SOLUTION:** *(c).* Since you are told by your friend that the most likely temperature is warm you would pack the warm weather suitcase.

2.3 Working with Distributions

Feeling slightly Bayesian, you contact the National Weather Service and obtain the distribution over temperatures plotted in Fig. 5.

1. Is the distribution in Fig. 5 consistent with the average temperature information and your friend's testimony? Explain.

   ★ **SOLUTION:** Yes. The average temperature is roughly $15°C$ and the most likely temperature is roughly $25°C$.

2. Given the utility function u(Temperature) and the distribution $\mathbf{P}$(Temperature) write down the integral you would need to evaluate to obtain the expected comfort.

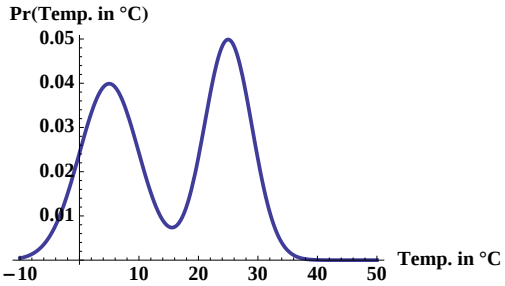


Figure 5: Temperature Distribution

★ **SOLUTION:** You want to maximize your expected utility with respect to your suitcase selection defined as:

$$\arg \max_{i \in \{a,b,c,d\}} \int u_i(T)\mathbf{P}(T)\, dT$$

3. Which suitcase maximizes your expected comfort and why?

★ **SOLUTION:** The suitcase that maximizes your comfort is clearly *(d)*, the suitcase for mixed conditions since it most closely matches the temperature distribution.

3 Linear Regression [85 Points]

The linear regression method is widely used in the medical domain. In this question you will work on a prostate cancer data from a study by Stamey et al.[1] You can download the data from http://www.cs.cmu.edu/~ggordon/10601/hws/hw3/prostatecancer.csv.

Your task is to predict the level of prostate-specific antigen (PSA) using a set of medical test results. PSA is a protein produced by the cells of the prostate gland. High levels of PSA often indicate the presence of prostate cancer or other prostate disorders.

The attributes are several clinical measurements on men who have prostate cancer. There are 8 attributes: log cancer volume `lcavol`, log prostate weight (`lweight`), log of the amount of benign prostatic hyperplasia (`lbph`), seminal vesicle invasion (`svi`), `age`, log of capsular penetration (`lcp`), Gleason score (`gleason`), and percent of Gleason scores of 4 or 5 (`pgg45`). `svi` and `gleason` are categorical, that is they take values either 1 or 0; others are real-valued. We will refer to these attributes as $A_1 = \text{lcavol}, A_2 = \text{lweight}, A_3 = \text{age}, A_4 = \text{lbph}, A_5 = \text{svi}, A_6 = \text{lcp}, A_7 = \text{gleason}, A_8 = \text{pgg45}$.

Each row of the input file describes one data point: the first column is the index of the data point, the following eight columns are attributes, and the tenth column gives the log PSA level `lpsa`, the response variable we are interested in. We already randomized the data and split it into three parts corresponding to training, validation and test sets. The last column of the file indicates whether the data point belongs to the training set, validation set or test set, indicated by '1' for training, '2' for validation and '3' for testing. The training data includes 57 examples; validation and test sets contain 20 examples each.

3.1 Inspecting the Data [10 points]

1. Calculate the correlation matrix of the 8 attributes and report it in a table. The table should be 8-by-8. You can use matlab functions.

[1] Stamey TA, Kabalin JN, McNeal JE et al. Prostate specific antigen in the diagnosis and treatment of the prostate. II. Radical prostatectomy treated patients. J Urol 1989;141:107683.

★ **SOLUTION:** See Table 2.

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa
lcavol	1.0000	0.2805	0.2250	0.0273	0.5388	0.6753	0.4324	0.4337	0.7345
lweight	0.2805	1.0000	0.3480	0.4423	0.1554	0.1645	0.0569	0.1074	0.4333
age	0.2250	0.3480	1.0000	0.3502	0.1177	0.1277	0.2689	0.2761	0.1696
lbph	0.0273	0.4423	0.3502	1.0000	-0.0858	-0.0070	0.0778	0.0785	0.1798
svi	0.5388	0.1554	0.1177	-0.0858	1.0000	0.6731	0.3204	0.4576	0.5662
lcp	0.6753	0.1645	0.1277	-0.0070	0.6731	1.0000	0.5148	0.6315	0.5488
gleason	0.4324	0.0569	0.2689	0.0778	0.3204	0.5148	1.0000	0.7519	0.3690
pgg45	0.4337	0.1074	0.2761	0.0785	0.4576	0.6315	0.7519	1.0000	0.4223
lpsa	0.7345	0.4333	0.1696	0.1798	0.5662	0.5488	0.3690	0.4223	1.0000

Table 2: Correlation coefficients of prostate cancer data. Answer to 3.1

2. Report the top 2 pairs of attributes that show the highest pairwise positive correlation and the top 2 pairs of attributes that show the highest pairwise negative correlation.

★ **SOLUTION:** The top 2 pairs that show the highest pairwise positive correlation are gleason-ppg4 (0.7519), lcavol-lcp(0.6731). Highest negative correlation, lbph-svi(-0.0858), lph-lcp(-0.0070).

3.2 Solving the Linear Regression Problem [45 points]

You will now try to find several models in order to predict the `lpsa` levels. The linear regression model is

$$Y = f(X) + \epsilon$$

where ϵ is a Gaussian noise variable and

$$f(X) = \sum_{j=0}^{p} w_j \phi_j(X)$$

where p is the number of basis functions (features), ϕ_j is the jth basis function, and w_j is the weight we wish to learn for the j^{th} basis function. In the models below, we will always assume that $\phi_0(X) = 1$ represents the intercept term.

1. Write a matlab function that takes the data matrix Φ and the column vector of responses y as an input and produces the least squares fit w as the output (refer to the lecture notes for the calculation of w).

★ **SOLUTION:** See below:

```
function what=lregress(Y,X)
% least square solution to linear regression
% X is the feature matrix
% Y is the response variable vector
what=inv(X'*X)*X'*Y;
end
```

2. You will create the following three models. Note that before solving each regression problem below, you should scale each feature vector to have a zero mean and unit variance. Don't forget to include the intercept column, $\phi_0(X) = 1$, after scaling the other features. Notice that since you shifted the attributes to have zero mean, in your solutions, the intercept term will be the mean of the response variable.

   - **Model1** Features are equal to input attributes, with the addition of a constant feature ϕ_0. That is, $\phi_0(X) = 1,\ \phi_1(X) = A_1,\ \ldots,\ \phi_8(X) = A_8$. Solve the linear regression problem and report the resulting feature weights. Discuss what it means for a feature to have a large negative weight, a large positive weight, or a small weight. Would you be able to comment on the weights, if you had not scaled the predictors to have the same variance? Report mean squared error (MSE) on the training and validation data.

     ★ **SOLUTION:** The weight vector for Model 1:
     $\vec{w} = [2.68265, 0.71796, 0.17843, -0.21235, 0.25752, 0.42998, -0.14179, 0.08745, 0.02928]$

     Features with a large positive and negative weight have big influence on the prediction. In the case of positive weights the higher they are the higher the response variable predicted, those features are positively correlated with the response variable. In the case negative weights the feature value is highly negatively correlated with the response variable. Weights close to zero does not have much affect on the prediction. If we had not scaled to the features variances to the same variance, the weights would have different scales, so would not be comparable and we would not be able to conclude features with each other in terms of their influence in predicting the response variable. Model 1's prediction MSE on training data is 0.4180, while it is 0.5005 on validation data. Notice the validation MSE is higher than the training error.

   - **Model2** Include additional features corresponding to pairwise products of the first six of the original attributes[2], i.e., $\phi_9(X) = A_1A_2,\ \ldots,\ \phi_{13}(X) = A_1A_6,\ \phi_{15}(X) = A_2A_3,\ \ldots,\ \phi_{23}(X) = A_5A_6$. First compute the features according to the formulas above using the unnormalized values, then shift and scale the new features to have zero mean and unit variance and add the column for the intercept term $\phi_0(X) = 1$. Report the five features whose weights achieved the largest absolute values.

     ★ **SOLUTION:** The largest five absolute values in descending order:
     lweight*age,lpbh,lweight,age,age*lpbh

   - **Model3** Starting with the results of Model1, drop the four features with the lowest weights (in absolute values). Build a new model using only the remaining features. Report the resulting weights.

     ★ **SOLUTION:** The features with have the lowest absolute weights in model 1 are: pgg45, gleason, lcp, lweight. The resulting weights:

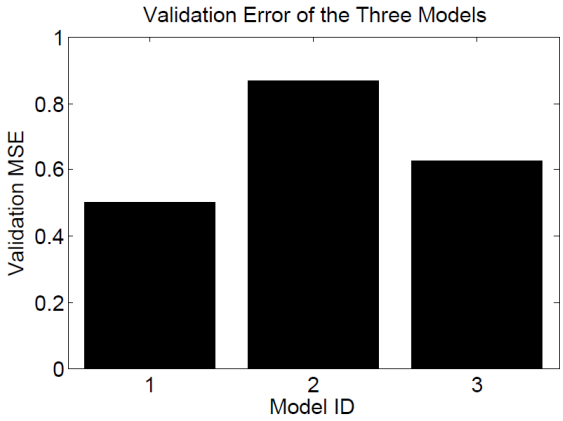
     $\vec{w} = [2.6827, 0.7164, -0.1735, 0.3441, 0.4095]$

3. Make two bar charts, the first to compare the *training* errors of the three models, the second to compare the *validation* errors of the three models. Which model achieves the best performance on the training data? Which model achieves the best performance on the validation data? Comment on differences between training and validation errors for individual models.

[2] These features are also called 'interactions', because they attempt to account for the effect of two attributes being simultaneously high or simultaneously low.

★ **SOLUTION:** See Fig. 6(a) and Fig. 6(b). Model2 achieves the best performance on the training data, whereas Model1 achieves the best performance on the validation data. Model2 suffer from overfitting indicated by the very good training model but low validation error. Model3 seems to be too simple, it has a higher training and a higher validation error compared to Model1. The features that are dropped are informative, as indicated by the lower training and validation errors.

(a) Training Error



(b) Validation Error

Figure 6: Training and validation errors for the three models.

4. Which of the models would you use for predicting the response variable? Explain.

★ **SOLUTION:** Model 1 since it achieves the best performance on the validation data. Model 2 overfits and Model 3 is too simple.

3.3 Ridge Regression [20 points]

For this question you will start with Model2 and employ regularization on it.

1. Write a matlab function to solve ridge regression. The function should take the data matrix Φ, the column vector of responses y, and the regularization parameter λ as the inputs and produce the least squares fit w as the output (refer to the lecture notes for the calculation of w). Do not penalize the intercept term. (You can achieve this by replacing the first column of the λI matrix with zeros.)

   ★ **SOLUTION:** See below:

```
function what=ridgeregress(Y,X,lambda)
% X is the feature matrix
% Y is the response vector
% what are the estimated weights
penal=lambda*eye(size(X,2));
penal(:,1)=0;
what=inv(X'*X+penal)*X'*Y;
end
```

2. You will create a plot exploring the effect of the regularization parameter on training and validation errors. The x-axis is the regularization parameter (on a log scale) and the y-axis is the mean squared error. Show two curves in the same graph, one for the training error and one for the validation error. Starting with $\lambda = 2^{-30}$, try 50 values: at each iteration increase λ by a factor of 2, so that for example the second iteration uses $\lambda = 2^{-29}$. For each λ, you need to train a new model.

   ★ **SOLUTION:** See Fig. 2.

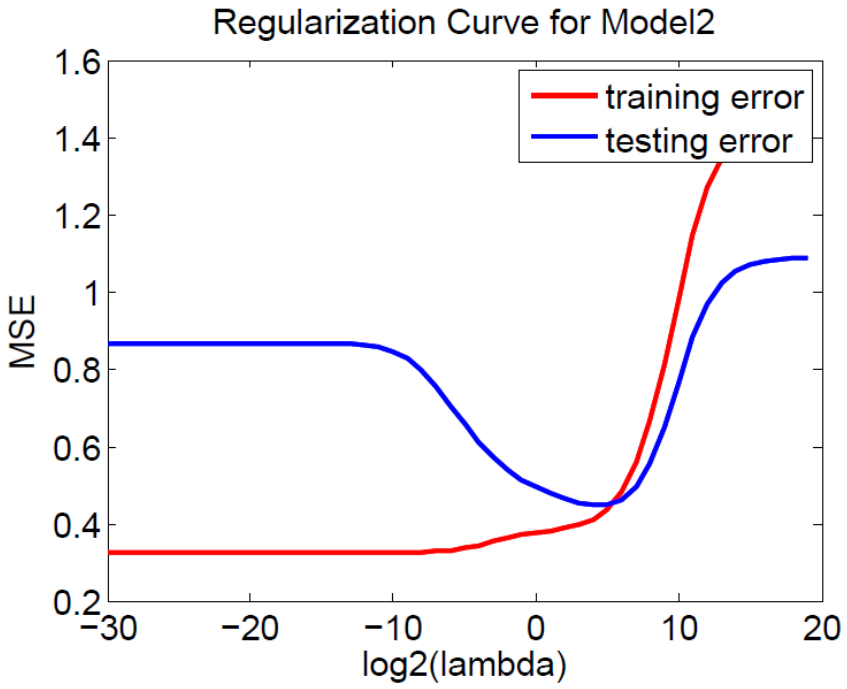


Figure 7: Training and validation errors with increasing regularization.

3. What happens to the training error as the regularization parameter increases? What about the validation error? Explain the curve in terms of overfitting, bias and variance.

★ **SOLUTION:** When the model is not regularized much (the left side of the graph), the training error is low and the validation error is high, indicating the model is too complex and overfitting to the training data. In that region bias is low and variance is high. As the regularization parameter increases, the bias increases and variance decreases. The overfitting problem is overcome as indicated by decreasing validation error and increasing training error. As regularization penalty increase too much, the model becomes getting too simple and start suffering from underfitting as can be shown by the poor performance on the training data.

4. What is the λ that achieves the lowest validation error and what is the validation error at that point? Compare this validation error to the Model2 validation error when no regularization was applied (you solved this in part 3.2). How does w differ in the regularized and unregularized versions, i.e., what effect did regularization have on the weights?

   ★ **SOLUTION:** $\log \lambda = 4, \lambda = 16$, achieves the lowest validation error, which is 0.447. This validation error is much less than the validation error of the model without regularization, which was 0.867. Regularized weights are smaller than unregularized weights. Regularization decreases the magnitude of the weights.

5. Is this validation error lower or higher than the validation error of the model you chose in 3.2.4? Which one should be your final model?

   ★ **SOLUTION:** The validation error of the penalized model($\lambda = 16$) is 0.447, which is lower than Model1's validation error. 0.5005. Therefore, this model is chosen.

3.4 Building the Final Model [10 points]

Now that you have decided on your model (features and possibly the regularization parameter), combine your training and validation data to make a combined training set, train your model on this combined training set, and calculate it on the test set. Report the training and test errors.

★ **SOLUTION:** The final models' training error is 0.40661 and test error is 0.58892