COGS 118A: ASSIGNMENT 3

Problem 1. Review of Bernoulli, binomial, and beta distributions (30)

(a) (5) Bishop problem 2.2:

[Normal]:

\[
\sum_{\{-1,1\}} p(x|\mu) = \left(\frac{1 - \mu}{2}\right)^{(1-(-1))/2} \left(\frac{1 + \mu}{2}\right)^{(1+(-1))/2} + \left(\frac{1 - \mu}{2}\right)^{(1-(-1))/2} \left(\frac{1 + \mu}{2}\right)^{(1+(1))/2}
\]

\[
= \frac{1 - \mu}{2} \cdot 1 + 1 \cdot \frac{1 + \mu}{2}
\]

\[
= \frac{1 + 1 - \mu + \mu}{2}
\]

\[
= 1
\]

[Mean]:

\[
\sum_{\{-1,1\}} xp(x|\mu) = -1 \cdot \left(\frac{1 - \mu}{2}\right)^{(1-(-1))/2} \left(\frac{1 + \mu}{2}\right)^{(1+(-1))/2} + 1 \cdot \left(\frac{1 - \mu}{2}\right)^{(1-(-1))/2} \left(\frac{1 + \mu}{2}\right)^{(1+(1))/2}
\]

\[
= \frac{1 - \mu}{2} + \frac{1 + \mu}{2}
\]

\[
= \frac{-1 + 1 + \mu}{2}
\]

\[
= \mu
\]

[Variance]:

\[
\mathbb{E}[x^2] = (-1)^2 \cdot \left(\frac{1 - \mu}{2}\right)^{(1-(-1))/2} \left(\frac{1 + \mu}{2}\right)^{(1+(-1))/2} + (1)^2 \cdot \left(\frac{1 - \mu}{2}\right)^{(1-(-1))/2} \left(\frac{1 + \mu}{2}\right)^{(1+(1))/2}
\]

\[
= \frac{1 - \mu}{2} + \frac{1 + \mu}{2}
\]

\[
= 1
\]

And so:

\[
\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = 1 - \mu^2
\]

[Entropy]:
\[ H(x) = - \sum_{\{-1,1\}} p(x) \log[p(x)] \]

\[ = - \sum_{\{-1,1\}} \left( \frac{1 - \mu}{2} \right)^{(1-x)/2} \left( \frac{1 + \mu}{2} \right)^{(1+x)/2} \log\left[ \frac{1 - \mu}{2} \right] \left( \frac{1 + \mu}{2} \right) \]

\[ = \left( \frac{1 - \mu}{2} \right) \log \left( \frac{1 - \mu}{2} \right) + \left( \frac{1 + \mu}{2} \right) \log \left( \frac{1 + \mu}{2} \right) \]

(b) (5) Bishop problem 2.4 (extra credit):

\[ \mathbb{E}[m] \]: Applying the product and chain rule to differentiate both sides we find that:

\[ \sum_{m=0}^{N} \binom{N}{m} m \mu^{m-1}(1 - \mu)^{N-m} - \mu^m (N - m)(1 - \mu)^{N-1-m} = 0 \]

By simple rearrangement:

\[ \sum_{m=0}^{N} \binom{N}{m} m \mu^{m-1}(1 - \mu)^{N-m} = \sum_{m=0}^{N} \binom{N}{m} \mu^m (N - m)(1 - \mu)^{N-1-m} \]

Multiplying both sides by \( \mu \) we find that:

\[ \mathbb{E}[m] = \sum_{m=0}^{N} \binom{N}{m} m \mu^m (1 - \mu)^{N-m} = \mu \sum_{m=0}^{N} \binom{N}{m} \mu^m (N - m)(1 - \mu)^{N-1-m} \]

We note that the \( m = N \) term of the sum is equal to zero, and thus:

\[ \mathbb{E}[m] = \mu \sum_{m=0}^{N-1} \binom{N}{m} \mu^m (N - m)(1 - \mu)^{N-1-m} \]

Then:

\[ \mathbb{E}[m] = \mu \sum_{m=0}^{N-1} \frac{N!}{(N - m)!m!} \mu^m (N - m)(1 - \mu)^{N-1-m} \]

\[ = N \mu \sum_{m=0}^{N-1} \frac{N-1!}{(N - 1 - m)!m!} \mu^m (1 - \mu)^{N-1-m} \]

\[ = N \mu \sum_{m=0}^{N-1} \binom{N-1}{m} \mu^m (1 - \mu)^{N-1-m} \]

\[ = N \mu \cdot 1 \quad (2.264) \]

[\text{var}[m]]: Left undone.

(c) (5) Bishop problem 2.6:

[mean]:
\[ E[\mu] = \int_0^1 \mu \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} d\mu \]
\[ = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a}(1-\mu)^{b-1} d\mu \]
\[ = \frac{\Gamma(a+b)}{a} \cdot \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \]  
\[
= \frac{\Gamma(a+b)}{a + b} \quad \text{Def of } \Gamma(x). 
\]

[var]:
\[ E[\mu^2] = \int_0^1 \mu^2 \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} d\mu \]
\[ = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 \mu^{a+1}(1-\mu)^{b-1} d\mu \]
\[ = \frac{\Gamma(a+b)}{a + b} \cdot \frac{\Gamma(a+b+2)}{(a+1)\Gamma(a)\Gamma(b)} \]  
\[
= \frac{\Gamma(a+b)}{(a+b+1)(a+b)} \]  

And so:
\[ \text{var}[x] = E[x^2] - E[x]^2 \]
\[ = \frac{a^2}{(a+b+1)(a+b)} - \frac{(a+b)^2}{(a^2+a)(a+b) - a^2(a+b+1)} \]
\[ = \frac{a^2}{(a+b+1)(a+b)^2} \]
\[ = \frac{ab}{(a+b+1)(a+b)^2} \]

as required.

[mode]:
\[ \frac{d}{d\mu} \text{Beta}(\lambda; a, b) = \frac{d}{d\mu} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1-\mu)^{b-1} \]
\[ = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{d}{d\mu} \mu^{a-1}(1-\mu)^{b-1} \]
\[ = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{d}{d\mu} \mu^{a-1}(1-\mu)^{b-1} \]
\[ = -\mu^{a-1}(b-1)(1-\mu)^{b-2} + (a-1)\mu^{a-2}(1-\mu)^{b-1} \]
\[ \Delta = 0 \]

Then:
\[ \mu^{a-1}(b-1)(1-\mu)^{b-2} = (a-1)\mu^{a-2}(1-\mu)^{b-1} \]
\[ \mu^{a-1-(a-2)}(1-\mu)^{b-2-(b-1)} = \frac{a-1}{b-1} \]
\[ \frac{\mu}{(1-\mu)} = \frac{a-1}{b-1} \]
\[ b\mu - \mu = a - \mu a - 1 + \mu \]
\[ \mu(a+b-2) = a - 1 \]
\[ \mu = \frac{a-1}{a+b-2} \]

(d) (5) The posterior will always be a beta distribution of form Beta(\(\lambda; a + a', b + b'\)) where a' and b' are the observed successes and failures.

The maximum likelihood estimate is the mode of the likelihood function. The maximum likelihood estimate together with the prior determine the mode of the posterior distribution; the influence of each is weighted by the number of real or notional data they contain.

The MAP estimate is the mode of this posterior distribution.

The Bayesian prediction is the mean of the posterior.

(e) (5) Bishop problem 2.7: Given the prior Beta(\(\lambda; a, b\)), we have:

Prior mean: \( \frac{a}{a+b} \) Posterior mean: \( \frac{a + m}{a + m + b + l} \) MLE: \( \frac{m}{m + l} \)

We reason that if there are no observations, then we should weight \( \lambda = 1 \). If \( m + l = a + b \) we should weight \( \lambda = 1/2 \). If \( m + l = 2(a + b) \) we should weight \( \lambda = 1/3 \). Then we attempt:

\[ \lambda = \frac{a + b}{a + b + m + l} \]

We find that:

\[ \frac{\lambda}{a + b} + (1 - \lambda) = \frac{m}{a + b + m + l} \]
\[ = \frac{a + b + m + l}{a + b + m + l} \cdot \frac{a + b + m + l}{a + b + m + l} \cdot \frac{m}{a + b + m + l} \]
\[ = \frac{a + b + m + l}{a + b + m + l} \]

as required.

(f) (5) Bishop problem 2.8: (extra credit) Left undone.

**Problem 2.** More on Gaussian distributions (30)

(a) (5) Bishop problem 2.13: (extra credit) Left undone.

(b) (5) Given:

\[ p(x) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp \left[ -\frac{1}{2}(x - \mu)^T\Sigma^{-1}(x - \mu) \right] \]
\[ H[x] = - \int p(x) \log(p(x)) \, dx \]
\[ = - \int p(x) \log\left( \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \right) \, dx \]
\[ = - \int p(x) \log\left( \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \right) \, dx - \int p(x) \log(\exp\left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right]) \, dx \]
\[ = - \log\left( \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \right) + \int p(x) \left[ \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \, dx \]

By \( \int p(x) \, dx = 1 \).

Noting that:
\[ \int p(x) \left[ \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \, dx = E \left[ \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \]

We reason by the symmetry of \( p(x) \) about \( \mu \) that:
\[ E \left[ \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] = 0 \]
\[ E \left[ \frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] = \frac{1}{2} \text{Tr} \left[ (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \]
\[ = \frac{1}{2} \text{Tr} \left[ (x - \mu)(x - \mu)^T \Sigma^{-1} \right] \]
\[ = \frac{1}{2} \text{Tr} \left[ \Sigma \Sigma^{-1} \right] \]
\[ = \frac{D}{2} \]

And so:
\[ H[x] = - \log\left( \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \right) + \frac{D}{2} \]
\[ = - \log\left( \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \right) + \frac{D}{2} \]
\[ = \log\left( \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \right) + \frac{D}{2} \]

(c) (5) Bishop problem 2.19 (hint: Bishop Appendix C): Left undone.

(d) (5) Bishop problem 2.21: We reason that the number of independent factors which determine a real, symmetric matrix as equal to the number of upper triangular positions in that matrix. For example, a \( 2 \times 2 \) matrix has the following three positions.
\[
\begin{bmatrix}
  a & b \\
  c & d
\end{bmatrix}
\]

We note that incrementing the size of the matrix to \( 3 \times 3 \) creates three additional independent factors:
\[
\begin{bmatrix}
  a & b & c \\
  c & d & e \\
  d & e & f
\end{bmatrix}
\]
And in general we can increment the size of an $n \times n$ matrix $A$ to construct an $n + 1 \times n + 1$ matrix $A'$ by attaching:

$$A' = \begin{bmatrix} A & B \\ C & d \end{bmatrix}$$

Where $B$ is $n \times 1$, $C$ is $1 \times n$, and $d$ is $1 \times 1$. Thus we have added $n + 1$ additional independent factors to those originally defining $A$. We proceed by induction.

Consider the case of $D = 1$. Then there is 1 independent factor defining our matrix. Then our condition is satisfied for $D = 1$ as $1(1 + 1)/2 = 1$.

Suppose that there exists some $k \in \mathbb{N}$ such that our condition holds for $D = k$. Then the number of independent factors in a $k \times k$ matrix is equal to $k(k + 1)/2$.

Now consider the case of $D = k + 1$. By our proof above, this is equal to:

$$\frac{k(k + 1)}{2} + (k + 1) = \frac{k^2 + k + 2k + 2}{2} = \frac{k^2 + 3k + 2}{2} = \frac{(k + 1)(k + 2)}{2}$$

as required, and thus our condition holds for $D = k + 1$. Then, by the principle of induction, our principle holds for all $D \in \mathbb{N}$.

(e) (5) Bishop problem 2.25:

Given the partitions:

$$x = \begin{pmatrix} x_a \\ x_b \\ x_c \end{pmatrix}, \mu = \begin{pmatrix} \mu_a \\ \mu_b \\ \mu_c \end{pmatrix}$$

$$\Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ba} & \Sigma_{bb} & \Sigma_{bc} \\ \Sigma_{ca} & \Sigma_{cb} & \Sigma_{cc} \end{bmatrix}$$

Reasoning that marginalizing out $x_c$ may be done by “slicing” $\Sigma_{ac}, \Sigma_{bc}, \Sigma_{cc}, \Sigma_{ca}$ and $\Sigma_{cb}$ from $\Sigma$ and $\mu_c$ from $\mu$ as formalized by (2.98), we find that $x_{ab}$, marginalizing out $x_c$, is normal and defined by:

$$\mu' = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$$

Then by (2.96), (2.97):

$$p(x_a|x_b) = \mathcal{N}(x|\mu_{a|b}, \Lambda_{aa}^{-1})$$

$$\mu_{a|b} = \mu_a - \Lambda_{aa}^{-1}\Lambda_{ab}(x_b - \mu_b)$$

(f) (5) Give an example of a 2-D covariance matrix, such that the components of the precision matrix aren’t equal to the inverse of the components of the covariance matrix.
\[
\begin{bmatrix}
4 & 2 \\
2 & 3
\end{bmatrix}^{-1} = \begin{bmatrix}
.375 & -.25 \\
-.25 & .5
\end{bmatrix}
\]

And yet \([4]^{-1} = .25 \neq .375\).

**Problem 3.** Matlab: Gaussian distribution (40)

```matlab
% Calculate the level curves, eigen values
p = 0:.01:2*pi;
s= sin(p); c= cos(p);
S1 = [1,5;5,1.5]; S2 = [1.5,0;0,1.5]; S3 = [1,0;0,1];
pl=1;
for j=1:3
[V,D]=eig(Sj);
D=diag(D);
subplot(3,3,pl); pl=pl+1; hold on grid on axis([-3,3,-3,3]);
for k=.5:1.5 clear x; for i =1:length(p) x(i,:)=c(i)*D(1)*V(1,:) + s(i)*D(2)*V(2,:); x(i,:)=k*x(i,:);
end h(1)=plot(x(:,1),x(:,2)); end
for k=1:2 h(2)=plot([0,V(1,k)],[0,V(2,k)],'r'); end
if j==2 legend(h,'Level Curves','Eigen Vectors'); end
end

% Calculate the marginal distributions.
x=[-3:.002:3];
for j=1:3 subplot(3,3,pl); pl=pl+1;
y=normpdf(x,0,Sj(1,1));
plot(x,y); if j==2 title('Marginal Distribution'); end axis([-3,3,0,.5]);
end

% Calculate the conditional distributions.
for j=1:3
L=inv(Sj);
mu=0-inv(L(1,1))*L(1,2)*(1-0) var=inv(L(1,1)); subplot(3,3,pl); pl=pl+1;
y=normpdf(x,mu,var);
plot(x,y); title(sprintf('p(x_{1} — x_{2}=1)
mean: %.2f, var: %.2f',mu,var)); axis([-3,3,0,.5]);
end
```