COGS 118A: ASSIGNMENT 5

Problem 1. Exponential family of distributions (36)

(a) (6) Bishop Eq. 2.226 gives us a general method for computing expected values of the (transformed) random variable $u(x)$ for the exponential family of distributions. It therefore gives us an alternative method for computing the moments of a distribution (by differentiating $\log g(\eta)$ rather than integrating $\int xp(x)dx$ or $\int x^2p(x)dx$ for instance, which may not be easy or even tractable). Use Eq. 226 to show that the the mean of the Bernoulli distribution (Eq. 2.196) is $\mu$. The standard representation (2.194) of the Bernoulli distribution is given in Bishop Eqs. 2.198-2.203.

(b) (6) Similar to part (a), use Eq. 226 to show that the mean and variance of a Gaussian distribution are indeed $\mu$ and $\sigma^2$. Recall that a Gaussian distribution can be written in the standard representation as a member of the exponential family according to Eqs. 2.220-2.223.

(c) (6) Recall that the conjugate prior of a Bernoulli distribution $\text{Bern}(x|\mu)$ is the beta distribution $\text{Beta}(\mu; a, b)$. Compare the beta distribution with the standard form of a conjugate prior for the exponential family (Eq. 2.229). What are $\chi$ and $\nu$ in this case?

(d) (6) Recall that the conjugate prior of a Gaussian distribution with unknown mean (and known variance), $\mathcal{N}(x|\mu; \sigma^2)$, is another Gaussian distribution $\mathcal{N}(\mu; \mu_0, \sigma_0^2)$. Compare it with the standard form of a conjugate prior (Eq. 2.229). What are $\chi$ and $\nu$ in this case?

(e) (6) Cast the beta distribution $\text{Beta}(\mu; a, b)$ into the standard representation (2.194) and show the following:

$$u(\mu) = \begin{bmatrix} \log \mu \\ \log(1 - \mu) \end{bmatrix} \quad \eta = \begin{bmatrix} a - 1 \\ b - 1 \end{bmatrix}$$

$$h(\mu) = 1 \quad g(\eta) = \frac{\Gamma(\eta_1 + \eta_2 + 2)}{\Gamma(\eta_1 + 1)\Gamma(\eta_2 + 1)}$$

(f) (6) Bishop 2.59

Problem 2. Bayesian model comparison (34)

Implicit in Bayesian inference is the assumption of a model class:

$$p(\eta|D, m) = \frac{p(D|\eta, m)p(\eta|m)}{p(D|m)}$$

where the term $p(D|m)$ is known as both the normalization constant and (model) evidence. It can be computed directly as an integral (or sum) of the numerator over all different settings of the model parameters $\eta$

$$p(D|m) = \int p(D|\eta, m)p(\eta|m)d\eta \quad (2-1)$$
Or, if there is a conjugate prior and therefore a simple posterior with known normalization constant, then it can be inferred analytically.

(a) (12) Suppose we have observations \( D = \{x_1, \ldots, x_n\} \) generated by a Gaussian with an unknown mean \( \mu \), and a variance \( \sigma^2 \) of either a small value \( \sigma^2_1 \) (model \( m_1 \)) or a large value \( \sigma^2_2 \) (model \( m_2 \)). According to Eq 2-1, we need to compute the following:

\[
p(D|\sigma^2_1) = \int p(D|\mu, \sigma^2_1)p(\mu)d\mu ,
\]

(2-2)

\[
p(D|\sigma^2_2) = \int p(D|\mu, \sigma^2_2)p(\mu)d\mu ,
\]

(2-3)

where we assumed that the mean is independent of the variance, \( p(\mu|\sigma^2) = p(\mu) \). We will first compute the evidence numerically in Matlab. Use \texttt{normrnd} to generate 50 samples from a Gaussian distribution with mean 0 and variance 1. We will consider the two hypotheses, \( m_1 : \sigma^2 = 1 \) and \( m_2 : \sigma^2 = 9 \). Discretize \( \mu \)-space as follows:

\[
\text{mu} = -5:.01:5;
\]

We assume that the prior over \( \mu \) is uniform (constant) for these values of \( \mu \) and 0 otherwise. The integrals in Eqs. 2-2 and 2-3 become sums:

\[
p(D|\sigma^2_1) = \sum_{\mu} p(D|\mu, \sigma^2_1)p(\mu) ,
\]

\[
p(D|\sigma^2_2) = \sum_{\mu} p(D|\mu, \sigma^2_2)p(\mu) .
\]

Use Matlab to compute the model evidences (use \texttt{normpdf} to compute the likelihoods). Now assume a uniform prior over the models, \( P(m_1) = P(m_2) = .5 \), what do you get for the posterior \( P(m_1|D) \)? (Make sure that your posterior distribution is normalized, i.e. \( P(m_1|D) + P(m_2|D) = 1 \).)

(b) (12) Now we will repeat (a) but use the fact that we have a conjugate prior for Gaussian to compute the model evidence analytically. We will assume a Gaussian prior with very large variance for \( \mu \), \( p(\mu) = \mathcal{N}(0, 100^2) \), essentially mimicking an (improper) uniform prior, as it is nearly flat in the region of interest. According to Bayes’ Rule,

\[
p(\mu|D; \sigma^2) = \frac{p(D|\mu; \sigma^2)p(\mu)}{p(D|\sigma^2)} .
\]

Because a Gaussian distribution over the mean is a conjugate prior for a Gaussian likelihood function, the posterior is also Gaussian, and we know exactly what the posterior mean and variance are (Bishop 2.141-2.143). We therefore know what the normalization constant is, which is the evidence. Use these facts to compute \( p(m_1|D) \) analytically, where \( D \) is the set of 50 data points from part (a). Again, assume uniform prior over the models, \( P(m_1) = P(m_2) = .5 \).

(c) (10) Compare the model posteriors obtained in (a) and (b). How do they differ? Which is more “correct”? Identify at least two possible reasons why they differ. Propose potential ways to mitigate those differences and discuss constraints on those proposals (i.e. what factors limit our ability to make the numerical solution as close to the analytical solution as possible?).
Problem 3. Mixture of Gaussians (30)

(a) (5) We will first generate a data set from a mixture of Gaussian distributions in Matlab. Suppose there are two Gaussian components, one with mean $\mu_1$, the other with mean $\mu_2$, and they share variance $\sigma_1^2 = \sigma_2^2$. Also assume the mixing proportions are $\pi = [\pi_1, \pi_2]^T$. In other words, this is the generative model:

$$
\begin{align*}
p(x|m_1; \mu_1, \sigma_1^2) &= \mathcal{N}(x; \mu_1, \sigma_1^2) \\
p(x|m_2; \mu_2, \sigma_2^2) &= \mathcal{N}(x; \mu_2, \sigma_2^2) \\
p(m_1; \pi) &= \pi_1 \\
p(m_2; \pi) &= \pi_2 \\
p(x; \mu_1, \sigma_1^2, \mu_2, \sigma_2^2, \pi) &= \pi_1 \mathcal{N}(x; \mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(x; \mu_2, \sigma_2^2)
\end{align*}
$$

Create a data set of 50 data points from this generative model with the following “true” parameters: $\mu_1^* = -1$, $\mu_2^* = 1$, $(\sigma_1^2)^* = (\sigma_2^2)^* = .5^2$, $\pi_1^* = \pi_2^* = .5$. Use hist to plot a histogram of the data. Label the axes.

(b) (15) Based on the data created in part (a), we will perform maximum likelihood estimation of the means of the Gaussians, assuming that we know the variances $\sigma_1^2 = \sigma_2^2 = .5^2$, and mixing proportions $\pi_1 = \pi_2 = .5$. Discretize possible values of $\mu_1$ and $\mu_2$ as follows:

$$
\begin{align*}
\text{mu1} &= -2:.01:2; \\
\text{mu2} &= -2:.01:2;
\end{align*}
$$

For each possible combination of values for $\mu_1$ and $\mu_2$, use part (a) to evaluate the likelihood of the entire dataset:

$$
p(D|\mu_1, \mu_2; \sigma_1^2, \sigma_2^2, \pi) = \prod_i p(x_i|\mu_1, \mu_2; \sigma_1^2, \sigma_2^2, \pi)
$$

and save the results into a matrix $A$. Use mesh(mu1,mu2,A) to visualize the 2-D landscape of the likelihood function. Why are there two peaks? Why is the likelihood function symmetric across the diagonal? Where are the peaks? How do they compare with the true values of $\mu_1^*$ and $\mu_2^*$? Discuss how the discrepancy may be reduced between the peaks and the true values of $\mu_1^*$ and $\mu_2^*$.

(a) (10) Describe a real-world data set which you believe could be modelled using a mixture of Gaussians. Argue why a mixture model is a sensible model for your real world data set. What do you expect the mixture components to represent? How many components (or clusters) do you think there should be?