

CS 281B/Stat 241B Homework Assignment 3 (due Apr. 21)

1. We have seen several examples in which the posterior expectation of a parameter is a convex combination of the prior mean and the maximum likelihood estimate. Show that this result holds in general for exponential family distributions. That is, compute the posterior expectation of the mean parameter (the mean parameter is the expectation of the sufficient statistic) under a conjugate prior and show that it has the claimed form.
2. In this problem we calculate the variance of the univariate Gaussian distribution in two different ways. (A third way would be the traditional method that involves computing an integral. Using the exponential family representation turns this integration into differentiation).
 - (a) Write the univariate Gaussian distribution in exponential family form and identify the sufficient statistic (a vector). Write down the relationship between the canonical parameters (η_1, η_2) and the parameters (μ, σ) .
 - (b) Write down the cumulant generating function $A(\eta_1, \eta_2)$.
 - (c) Compute the variance by the appropriate combination of first derivatives of A .
 - (d) Compute the variance by the appropriate combination of a second derivative of A .
3. Let $\theta = (\theta_1, \theta_2, \dots, \theta_k)$ be distributed according to a Dirichlet distribution, with parameter vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_k)$.
 - (a) Compute the expectation $E[\theta_i]$.
 - (b) Express the Dirichlet distribution in the exponential family form. Identify the sufficient statistic (a vector) and compute the expectation of the sufficient statistic.
4. In this problem you are to write out all of the conditionals needed to implement a Gibbs sampler for a Bayesian approach to Gaussian mixture modeling. You are to do this in the univariate setting. The model is:

$$p(x|\mu, \sigma, \pi) = \sum_{i=1}^k \pi_i N(x|\mu_i, \sigma_i),$$

where $N(x|\mu_i, \sigma_i)$ is a univariate Gaussian. Each of the parameters, μ , σ , and π , are endowed with conjugate priors (Gaussian, inverse gamma, and Dirichlet, respectively). (BONUS: Do it for the multivariate Gaussian instead).

5. Consider a regression problem: $y = f(x) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Assume that we have N IID observations, $\{(x_i, y_i)\}_{i=1}^N$. Recall that in the Gaussian process setting, the estimate of the regression curve takes the following form:

$$\hat{f}(x) = \mathbf{k}^T K^{-1} \mathbf{y}, \tag{1}$$

where the i th component of the vector \mathbf{k} takes the value $k(x, x_i)$ for a covariance kernel $k(\cdot, \cdot)$, where K is the matrix $K_{ij} = k(x_i, x_j) + \sigma^2 \delta(i, j)$, and where \mathbf{y} is the data vector (the i th component of \mathbf{y} is the observation y_i).

As we have seen in class, by grouping together K^{-1} and \mathbf{y} , we can write Eq. (1) in the following way:

$$\hat{f}(x) = \sum_{i=1}^N \alpha_i k(x, x_i), \quad (2)$$

where the coefficients α_i depend on both the values $\{x_i\}$ and the values $\{y_i\}$.

We can also group together \mathbf{k}^T and K^{-1} , and obtain a different expansion:

$$\hat{f}(x) = \sum_{j=1}^N b(x, x_j) y_j, \quad (3)$$

which is a weighted sum of the observations $\{y_i\}$, and where $b(x, x_j)$ does not depend on the $\{y_i\}$. The function $b(x, x_j)$ is a so-called *equivalent kernel*. Note that an “equivalent kernel” is *not* in general a kernel function in our sense; i.e., a positive definite function.

It is the goal of this problem to explore some of the relationships between these two expansions. We will do so in the context of the Gaussian process approach to regression.

- (a) Fit a Gaussian process regression to the data in the file `gp.dat` on the course website. These data were generated from $y = \sin(4x) + \epsilon$, with $\sigma^2 = 0.333$. Use a Gaussian covariance kernel:

$$k(x, x') = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left\{-\frac{1}{2\tau^2}(x - x')^2\right\}.$$

Set the variance parameter τ^2 equal to 0.12. Use this kernel to evaluate the fit in Eq. (1). (Don’t forget to add σ^2 to the diagonal when you form the matrix K). Plot the resulting fit. (I.e., evaluate $\hat{f}(x)$ at the points on a grid). Also include the data points in your plot, as well as a plot of the true regression curve $f(x) = \sin(4x)$.

- (b) Plot the kernel function $k(x, x_i)$, evaluated at the data point $x_i = 0.1210$ and evaluated at the data point $x_i = 0.5488$.
- (c) Plot the equivalent kernel function $b(x, x_i)$, evaluated at the data point $x_i = 0.1210$ and evaluated at the data point $x_i = 0.5488$.
- (d) Comment on the similarities and differences between these two types of “kernels.”
- (e) Now redo the problem using a different covariance kernel. In particular, use the *multiquadric kernel*:

$$k(x, x') = \sqrt{\tau^2 + (x - x')^2}.$$

Set the parameter τ^2 equal to 0.08.