Practice Problem Set I

Bayesian Estimation, Conditional Models, Gaussian Process

(From the text books by Kevin Murphy, and David Barber)

Exercise 3.4 Beta updating from censored likelihood

(Source: Gelman.) Suppose we toss a coin $n = 5$ times. Let X be the number of heads. We observe that there are fewer than 3 heads, but we don't know exactly how many. Let the prior probability of heads be $p(\theta) = \text{Beta}(\theta|1, 1)$. Compute the posterior $p(\theta|X < 3)$ up to normalization constants, i.e., derive an expression proportional to $p(\theta, X < 3)$. Hint: the answer is a mixture distribution.

Exercise 3.5 Uninformative prior for log-odds ratio

Let

$$\phi = \text{logit}(\theta) = \log \frac{\theta}{1-\theta} \tag{3.91}$$

Show that if $p(\phi) \propto 1$, then $p(\theta) \propto \text{Beta}(\theta|0, 0)$. Hint: use the change of variables formula.

Exercise 3.6 MLE for the Poisson distribution

The Poisson pmf is defined as $\text{Poi}(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$, for $x \in \{0, 1, 2, \ldots\}$ where $\lambda > 0$ is the rate parameter. Derive the MLE.

Exercise 3.8 MLE for the uniform distribution

(Source: Kaelbling.) Consider a uniform distribution centered on 0 with width $2a$. The density function is given by

$$p(x) = \frac{1}{2a} I(x \in [-a, a]) \tag{3.92}$$

a. Given a data set $x_1, \ldots, x_n$, what is the maximum likelihood estimate of a (call it $\hat{a}$)?

b. What probability would the model assign to a new data point x_{n+1} using $\hat{a}$?

c. Do you see any problem with the above approach? Briefly suggest (in words) a better approach.

Exercise 3.14 Posterior predictive for Dirichlet-multinomial

(Source: Koller.).

a. Suppose we compute the empirical distribution over letters of the Roman alphabet plus the space character (a distribution over 27 values) from 2000 samples. Suppose we see the letter "e" 260 times. What is $p(x_{2001} = e|\mathcal{D})$, if we assume $\boldsymbol{\theta} \sim \text{Dir}(\alpha_1, \ldots, \alpha_{27})$, where $\alpha_k = 10$ for all k?

b. Suppose, in the 2000 samples, we saw "e" 260 times, "a" 100 times, and "p" 87 times. What is $p(x_{2001} = p, x_{2002} = a|\mathcal{D})$, if we assume $\boldsymbol{\theta} \sim \text{Dir}(\alpha_1, \ldots, \alpha_{27})$, where $\alpha_k = 10$ for all k? Show your work.

Exercise 3.10 Taxicab (tramcar) problem

Suppose you arrive in a new city and see a taxi numbered 100. How many taxis are there in this city? Let us assume taxis are numbered sequentially as integers starting from 0, up to some unknown upper bound θ. (We number taxis from 0 for simplicity; we can also count from 1 without changing the analysis.) Hence the likelihood function is $p(x) = U(0, \theta)$, the uniform distribution. The goal is to estimate θ. We will use the Bayesian analysis from Exercise 3.9.

a. Suppose we see one taxi numbered 100, so $\mathcal{D} = \{100\}$, $m = 100$, $N = 1$. Using an (improper) non-informative prior on θ of the form $p(\theta) = Pa(\theta|0, 0) \propto 1/\theta$, what is the posterior $p(\theta|D)$?

b. Compute the posterior mean, mode and median number of taxis in the city, if such quantities exist.

c. Rather than trying to compute a point estimate of the number of taxis, we can compute the predictive density over the next taxicab number using

$$p(D'|D, \alpha) = \int p(D'|\theta)p(\theta|D, \alpha)d\theta = p(D'|\beta) \tag{3.96}$$

where $\alpha = (b, K)$ are the hyper-parameters, $\beta = (c, N + K)$ are the updated hyper-parameters. Now consider the case $D = \{m\}$, and $D' = \{x\}$. Using Equation 3.95, write down an expression for

$$p(x|D, \alpha) \tag{3.97}$$

As above, use a non-informative prior $b = K = 0$.

d. Use the predictive density formula to compute the probability that the next taxi you will see (say, the next day) has number 100, 50 or 150, i.e., compute $p(x = 100|D, \alpha)$, $p(x = 50|D, \alpha)$, $p(x = 150|D, \alpha)$.

Exercise 4.14 MAP estimation for 1D Gaussians

(Source: Jaakkola.)

Consider samples $x_1, \ldots, x_n$ from a Gaussian random variable with known variance σ^2 and unknown mean μ. We further assume a prior distribution (also Gaussian) over the mean, $\mu \sim \mathcal{N}(m, s^2)$, with fixed mean m and fixed variance s^2. Thus the only unknown is μ.

a. Calculate the MAP estimate $\hat{\mu}_{MAP}$. You can state the result without proof. Alternatively, with a lot more work, you can compute derivatives of the log posterior, set to zero and solve.

b. Show that as the number of samples n increase, the MAP estimate converges to the maximum likelihood estimate.

c. Suppose n is small and fixed. What does the MAP estimator converge to if we increase the prior variance s^2?

d. Suppose n is small and fixed. What does the MAP estimator converge to if we decrease the prior variance s^2?

Exercise 7.6 MLE for simple linear regression

Simple linear regression refers to the case where the input is scalar, so $D = 1$. Show that the MLE in this case is given by the following equations, which may be familiar from basic statistics classes:

$$w_1 = \frac{\sum_i (x_i - \overline{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i x_i y_i - N\overline{x}\,\overline{y}}{\sum_i x_i^2 - N\overline{x}^2} \approx \frac{\text{cov}[X, Y]}{\text{var}[X]} \tag{7.99}$$

$$w_0 = \bar{y} - w_1\bar{x} \approx \mathbb{E}[Y] - w_1\mathbb{E}[X] \tag{7.100}$$

Exercise 8.5 Symmetric version of ℓ_2 regularized multinomial logistic regression

(Source: Ex 18.3 of (Hastie et al. 2009).)

Multiclass logistic regression has the form

$$p(y = c|\mathbf{x}, \mathbf{W}) = \frac{\exp(w_{c0} + \mathbf{w}_c^T \mathbf{x})}{\sum_{k=1}^{C} \exp(w_{k0} + \mathbf{w}_k^T \mathbf{x})} \tag{8.128}$$

where $\mathbf{W}$ is a $(D+1) \times C$ weight matrix. We can arbitrarily define $\mathbf{w}_c = \mathbf{0}$ for one of the classes, say $c = C$, since $p(y = C|\mathbf{x}, \mathbf{W}) = 1 - \sum_{c=1}^{C-1} p(y = c|\mathbf{x}, \mathbf{w})$. In this case, the model has the form

$$p(y = c|\mathbf{x}, \mathbf{W}) = \frac{\exp(w_{c0} + \mathbf{w}_c^T \mathbf{x})}{1 + \sum_{k=1}^{C-1} \exp(w_{k0} + \mathbf{w}_k^T \mathbf{x})} \tag{8.129}$$

If we don't "clamp" one of the vectors to some constant value, the parameters will be unidentifiable. However, suppose we don't clamp $\mathbf{w}_c = \mathbf{0}$, so we are using Equation 8.128, but we add ℓ_2 regularization by optimizing

$$\sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i, \mathbf{W}) - \lambda \sum_{c=1}^{C} ||\mathbf{w}_c||_2^2 \tag{8.130}$$

Show that at the optimum we have $\sum_{c=1}^{C} \hat{w}_{cj} = 0$ for $j = 1 : D$. (For the unregularized $\hat{w}_{c0}$ terms, we still need to enforce that $w_{0C} = 0$ to ensure identifiability of the offset.)

Exercise 9.2 The MVN is in the exponential family

Show that we can write the MVN in exponential family form.

Exercise 185. *The exercise concerns Bayesian regression.*

1. *Show that for*

$$f = \mathbf{w}^{\mathsf{T}} \mathbf{x} \tag{18.4.1}$$

and $p(\mathbf{w}) \sim \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{\Sigma})$, that $p(f|\mathbf{x})$ is Gaussian distributed. Furthermore, find the mean and covariance of this Gaussian.

2. *Consider a target point $t = f + \epsilon$, where $\epsilon \sim \mathcal{N}(\epsilon|0, \sigma^2)$. What is $p(f|t, \mathbf{x})$?*

Exercise 191. *Show that the sample covariance matrix with elements $S_{ij} = \sum_{n=1}^{N} x_i^n x_j^n / N - \bar{x}_i \bar{x}_j$, where $\bar{x}_i = \sum_{n=1}^{N} x_i^n / N$, is positive semidefinite.*

Exercise 192. *Show that*

$$k(x - x') = e^{-|\sin(x - x')|} \tag{19.8.1}$$

is a covariance function.

Exercise 186. *A Bayesian Linear Parameter regression model is given by*

$$y^n = \mathbf{w}^{\mathsf{T}}\boldsymbol{\phi}(\mathbf{x}^n) + \eta^n \tag{18.4.2}$$

In vector notation $\mathbf{y} = \left(y^1, \dots, y^N\right)^{\mathsf{T}}$ *this can be written*

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\eta} \tag{18.4.3}$$

with $\boldsymbol{\Phi}^{\mathsf{T}} = \left[\boldsymbol{\phi}(\mathbf{x}^1), \dots, \boldsymbol{\phi}(\mathbf{x}^N)\right]$ *and* $\boldsymbol{\eta}$ *is a zero mean Gaussian distributed vector with covariance* $\beta^{-1}\mathbf{I}$*. An expression for the marginal likelihood of a dataset is given in equation (18.1.19). A more compact expression can be obtained by considering*

$$p(y^1, \dots, y^N | \mathbf{x}^1, \dots, \mathbf{x}^N, \Gamma) \tag{18.4.4}$$

Since y^n *is linearly related to* $\mathbf{x}^n$ *through* $\mathbf{w}$*. Then* $\mathbf{y}$ *is Gaussian distributed with mean*

$$\langle \mathbf{y} \rangle = \boldsymbol{\Phi} \langle \mathbf{w} \rangle = \mathbf{0} \tag{18.4.5}$$

and covariance matrix

$$\left\langle \mathbf{y}\mathbf{y}^{\mathsf{T}} \right\rangle - \langle \mathbf{y} \rangle \langle \mathbf{y} \rangle^{\mathsf{T}} = \left\langle (\boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\eta})(\boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\eta})^{\mathsf{T}} \right\rangle \tag{18.4.6}$$

For $p(\mathbf{w}) = \mathcal{N}\left(\mathbf{w} | \mathbf{0}, \alpha^{-1}\mathbf{I}\right)$*:*

1. *Show that the covariance matrix can be expressed as*

$$\mathbf{C} = \frac{1}{\beta}\mathbf{I} + \frac{1}{\alpha}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathsf{T}} \tag{18.4.7}$$

2. *Hence show that the log marginal likelihood can we written as*

$$\log p(y^1, \dots, y^N | \mathbf{x}^1, \dots, \mathbf{x}^N, \Gamma) = -\frac{1}{2}\log\det(2\pi\mathbf{C}) - \frac{1}{2}\mathbf{y}^{\mathsf{T}}\mathbf{C}^{-1}\mathbf{y} \tag{18.4.8}$$

Exercise 195. *For a covariance function*

$$k_1(\mathbf{x}, \mathbf{x}') = f\left(\left(\mathbf{x} - \mathbf{x}'\right)^{\mathsf{T}}\left(\mathbf{x} - \mathbf{x}'\right)\right) \tag{19.8.5}$$

show that

$$k_2(\mathbf{x}, \mathbf{x}') = f\left(\left(\mathbf{x} - \mathbf{x}'\right)^{\mathsf{T}}\mathbf{A}\left(\mathbf{x} - \mathbf{x}'\right)\right) \tag{19.8.6}$$

is also a valid covariance function for a positive definite symmetric matrix $\mathbf{A}$*.*

Exercise 197 (Vector regression)**.** *Consider predicting a vector output* $\mathbf{y}$ *given training data* $\mathcal{X} \cup \mathcal{Y} = \{\mathbf{x}^n, \mathbf{y}^n, n = 1, \dots, n\}$*. To make a GP predictor*

$$p(\mathbf{y}^* | \mathbf{x}^*, \mathcal{X}, \mathcal{Y}) \tag{19.8.8}$$

we need a Gaussian model

$$p(\mathbf{y}^1, \dots, \mathbf{y}^N, \mathbf{y}^* | \mathbf{x}^1, \dots, \mathbf{x}^n, \mathbf{x}^*) \tag{19.8.9}$$

A GP requires then a specification of the covariance $c(y_i^m, y_j^n | \mathbf{x}^n, \mathbf{x}^m)$ *of the components of the outputs for two different input vectors. Show that under the dimension independence assumption*

$$c(y_i^m, y_j^n | \mathbf{x}^n, \mathbf{x}^m) = c_i(y_i^m, y_i^n | \mathbf{x}^n, \mathbf{x}^m)\delta_{ij} \tag{19.8.10}$$

where $c_i(y_i^m, y_i^n | \mathbf{x}^n, \mathbf{x}^m)$ *is a covariance function for the* i^{th} *dimension, that separate GP predictors can be constructed independently, one for each output dimension* i*.*