CPSC 540 Assignment 7 (due November 26)

Bayesian Learning

Please put your name and student number on the assignment, there are also potential bonus marks for the submission format:

- +2 point if the submission is done in LaTeX.
- +1 point if hte submission is typed.
- no bonus points if the submission is hard-written.

Keep in mind that only the top 6 assignments count, so if you are happy with your mark on the first six assignments then you do not have to do this assignment.

1 Bayes Baysics

Consider a $y \in \{1, 2, 3\}$ following a multinoulli distribution with parameters $\theta = \{\theta_1, \theta_2, \theta_3\}$,

$$y|\theta \sim \text{Mult}(\theta_1, \theta_2, \theta_3).$$

We'll assume that θ follows a Dirichlet distribution (the conjugate prior to the multinoulli) with parameters $\alpha = \{\alpha_1, \alpha_2, \alpha_3\}$,

$$\theta \sim \mathcal{D}(\alpha_1, \alpha_2, \alpha_3).$$

Thus we have

$$p(y|\theta, \alpha) = p(y|\theta) = \theta_1^{I(y=1)}\theta_2^{I(y=2)}\theta_3^{I(y=3)}, \quad p(\theta|\alpha) = \frac{\Gamma(\alpha_1 + \alpha_2 + \alpha_3)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}\theta_1^{\alpha_1 - 1}\theta_2^{\alpha_2 - 1}\theta_3^{\alpha_3 - 1}.$$

1.1 Posterior Distribution

Derive the posterior distribution,

$$p(\theta|y, \alpha).$$

MLE Estimate

The MLE estimate for θ is the solution to

$$\max_\theta \log p(y|\theta), \text{ subject to } \theta_1 + \theta_2 + \theta_3 = 1,$$

and that all the $\theta_i \geq 0$. It turns out we can ignore the bound constraints, so to compute the MLE we need to solve an equality-constrained problem. To solve a problem of the form

$$\min_x f(x) \text{ subject to } Ax = b,$$

you need to find a stationary point of the Lagrangian function,

$$L(x, z) = f(x) + z^T(Ax - b).$$

For the MLE problem, we have

$$L(y, \lambda) = I(y = 1) \log \theta_1 + I(y = 2) \log \theta_2 + I(y = 3) \log \theta_3 + \lambda(1 - (\theta_1 + \theta_2 + \theta_3)) + (\text{constant}).$$

Taking the gradient we get

$$\nabla L(y, \lambda) = \begin{bmatrix} \frac{I(y=1)}{\theta_1} - \lambda \\ \frac{I(y=2)}{\theta_2} - \lambda \\ \frac{I(y=3)}{\theta_3} - \lambda \\ 1 - (\theta_1 + \theta_2 + \theta_3) \end{bmatrix}$$

To make this gradient equal to zero, we need $(\theta_1 + \theta_2 + \theta_3) = 1$ and we also need

$$\lambda\theta_i = I(y = i),$$

for all i. Summing these constraints over i we get

$$\sum_{i=1}^{3} \lambda\theta_i = \sum_{i=1}^{3} I(y = i),$$

or equivalently that

$$\lambda \sum_{i=1}^{3} \theta_i = \sum_{i=1}^{3} I(y = i).$$

and thus that

$$\lambda = 1.$$

Plugging this value of λ into the gradient of the Lagrangian and setting it to zero we get

$$\theta_i = \frac{I(y = i)}{1}.$$

Thus the MLE will give $\theta_i = 1$ for the even that actually happened, and $\theta_i = 0$ for the other two events.

1.2 MAP Estimate

Use the reasoning above to compute the MAP estimate for θ,

$$\max_\theta p(\theta | y, \alpha).$$

1.3 Marginal Likelihood

Derive the marginal likelihood of y given the hyper-parameters α,

$$p(y|\alpha) = \int p(y, \theta|\alpha)d\theta,$$

Hint: Because $\int p(\theta|\alpha)d\theta = 1$, we know that $\int \theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\theta_3^{\alpha_3-1}d\theta = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1+\alpha_2+\alpha_3)}$. You can use $D(\alpha) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}{\Gamma(\alpha_1+\alpha_2+\alpha_3)}$ to represent the normalizing constant of the prior and $D(\alpha^+)$ to give the normalizing constant of the posterior.

1.4 Posterior Mean

Compute the posterior mean estimate for θ,

$$\mathbb{E}_{\theta|y,\alpha}[\theta_i] = \int \theta_i p(\theta|y, \alpha)d\theta,$$

which (after some manipulation) should not involve any Γ functions.

Hint: You will also need to use that $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. You may find it a bit cleaner to parameterize the posterior in terms of $\beta_j = I(y = j) + \alpha_j$, and convert back once you have the final result.

1.5 Posterior Predictive

Derive the posterior predictive distribution for a new independent observation $\hat{y}$ given y,

$$p(\hat{y}|y, \alpha) = \int p(\hat{y}, \theta|y, \alpha)d\theta.$$

2 Marginal Likelihood

The function *basisDemo* gives a solution to Question 5 of Assignment 2. Assume the following Bayesian linear regression model of the data and parameters,

$$y_i \sim \mathcal{N}(w^T\phi(x_i), \sigma^2), \quad w_j \sim \mathcal{N}(0, \lambda^{-1}).$$

The marginal likelihood of the target vector y given the design matrix X under this model is given by

$$p(y|X, \sigma, \lambda) = (2\pi)^{-N/2}|C|^{-1/2}\exp\left(-\frac{y'C^{-1}y}{2}\right),$$

where

$$C = \sigma^2 I + \lambda\Phi(X)\Phi(X)^T.$$

Add the calculation of the logarithm of the marginal likelihood, $\log p(y|X, \sigma, \lambda)$, to the inner loop of this demo. You can use the function *logdet* to compute the logarithm of the determinant, $\log|C|$. Hand in the code to compute the marginal likelihood, and report the order that optimizes the marginal likelihood for $\lambda = 1$ and σ^2 set to 1, 10, and 0.1.

3 Type II Maximum Likelihood

The function *regPathDemo* gives a solution to Question 5 of Assignment 3. An alternative to ℓ_1-regularization for achieving sparsity is via type II maximum likelihood. Consider the model

$$y_i \sim \mathcal{N}(w^T \phi(x_i), \sigma^2), \quad w_j \sim \mathcal{N}(0, \lambda_j^{-1}).$$

Automatic relevance determination (ARD) corresponds to Type II maximum likelihood (maximum likelihood) estimation of the λ_j in this model. The marginal likelihood is

$$p(y_i|x_i, \sigma, \lambda) = (2\pi)^{-N/2} |C|^{-1/2} \exp\left(-\frac{y' C^{-1} y}{2}\right),$$

where

$$C = \sigma^2 I + \Phi(X) \Lambda \Phi(X)^T,$$

and Λ is a diagonal matrix containing the λ_j on the diagonal.

The first algorithm for this problem is known as MacKay's method (Section 13.7.4.2 of MLAPP). It uses the update

$$\lambda_j \leftarrow \frac{w_j^2}{1 - \frac{V_{j,j}}{\lambda_j}},$$

where

$$w \leftarrow \Lambda X^T C^{-1} y, \quad V \leftarrow \Lambda - \Lambda X^T C^{-1} X \Lambda.$$

Implement MacKay's method and modify the demo so that it plots the posterior means w above as σ^2 is varied (over the same range that λ is currently varied for the ℓ_1-regularization approach). (You will have to do something to avoid dividing by 0, one common strategy is to set λ_j to some small value when this occurs.) Hand in the *ARD* function and the updated plot.