EE5907/EE5027 Week 6: Bayesian Statistics

The following questions are from Kevin Murphy's (KM) book "Machine Learning: A Probabilistic Perspective".

Exercise 5.1 Proof that a mixture of conjugate priors is indeed conjugate

Consider the mixture of conjugate priors: $p(\theta) = \sum_k p(z = k)p(\theta|z = k)$ Derive the following equation:

$$p(\theta|\mathcal{D}) = \sum_k p(z = k|\mathcal{D})p(\theta|\mathcal{D}, z = k)$$

Exercise 5.3 Reject option in classifiers

In many classification problems one has the option either of assigning x to class j or, if you are too uncertain, of choosing the **reject option**. If the cost for rejects is less than the cost of falsely classifying the object, it may be the optimal action. Let α_i mean you choose action i, for $i = 1 : C + 1$, where C is the number of classes and $C + 1$ is the reject action. Let $Y = j$ be the true (but unknown) **state of nature**. Define the loss function as follows

$$\lambda(\alpha_i|Y = j) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, \cdots, C\} \\ \lambda_r & \text{if } i = C + 1 \\ \lambda_s & \text{otherwise} \end{cases} \tag{1}$$

In otherwords, you incur 0 loss if you correctly classify, you incur λ_r loss (cost) if you choose the reject option, and you incur λ_s loss (cost) if you make a substitution error (misclassification).

a. Show that the minimum risk is obtained if we decide $Y = j$ if $p(Y = j|x) \geq p(Y = k|x)$ for all k (i.e., j is the most probable class; $1 \leq j, k \leq C$) and if $p(Y = j|x) \geq 1 - \frac{\lambda_r}{\lambda_s}$; otherwise we decide to reject.

b. Describe qualitatively what happens as λ_r/λ_s is increased from 0 to 1 (i.e., the relative cost of rejection increases).

Exercise 5.7 Bayes model averaging helps predictive accuracy

Let Δ be a quantity that we want to predict, let $\mathcal{D}$ be the observed data and $\mathcal{M}$ be a finite set of models. Suppose our action is to provide a probabilistic prediction $p()$, and the loss function is $L(\Delta, p()) = -\log p(\Delta)$. We can either perform Bayes model averaging and predict using

$$p^{BMA}(\Delta) = \sum_{m \in \mathcal{M}} p(\Delta|m, \mathcal{D}) p(m|\mathcal{D}) \tag{2}$$

or we could predict using any single model m' (a plugin approximation)

$$p^{M}(\Delta) = p(\Delta|m', \mathcal{D}) \tag{3}$$

Show that, for all models $m \in \mathcal{M}$, the posterior expected loss using BMA is lower, i.e.,

$$\mathbb{E}\left[L(\Delta, p^{BMA})\right] \leq \mathbb{E}[L(\Delta, p^{M})] \tag{4}$$

where the expectation over Δ is with respect to

$$p(\Delta|\mathcal{D}) = \sum_{m \in \mathcal{M}} p(\Delta|m, \mathcal{D}) p(m|\mathcal{D}) \tag{5}$$

Hint: use the non-negativity of the KL divergence.

Exercise 5.8 MLE and model selection for a 2d discrete distribution

Let $x \in \{0, 1\}$ denote the result of a coin toss ($x = 0$ for tails, $x = 1$ for heads). The coin is potentially biased, so that heads occurs with probability θ_1. Suppose that someone else observes the coin flip and reports to you the outcome, y. But this person is unreliable and only reports the result correctly with probability θ_2; i.e., $p(y|x, \theta_2)$ is given by

	$y = 0$	$y = 1$
$x = 0$	θ_2	$1 - \theta_2$
$x = 1$	$1 - \theta_2$	θ_2

Assume that θ_2 is independent of x and θ_1.

a. Write down the joint probability distribution $p(x, y|\boldsymbol{\theta})$ as a 2×2 table, in terms of $\boldsymbol{\theta} = (\theta_1, \theta_2)$.

b. Suppose have the following dataset: $\boldsymbol{x}$ = (1,1,0,1,1,0,0), $\boldsymbol{y}$ = (1,0,0,0,1,0,1). What are the MLEs for θ_1 and θ_2? Justify your answer. Hint: note that the likelihood function factorizes,

$$p(x, y|\theta) = p(y|x, \theta_2) p(x|\theta_1) \tag{6}$$

What is $p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_2)$ where M_2 denotes this 2-parameter model? (You may leave your answer in fractional form if you wish.)

c. Now consider a model with 4 parameters, $\boldsymbol{\theta} = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})$, representing $p(x, y|\theta) = \theta_{x,y}$. (Only 3 of these parameters are free to vary, since they must sum to one.) What is the MLE of $\boldsymbol{\theta}$? What is $p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_4)$ where M_4 denotes this 4-parameter model?

d. Suppose we are not sure which model is correct. We compute the leave-one-out cross validated log likelihood of the 2-parameter model and the 4-parameter model as follows:

$$L(m) = \sum_{i=1}^{n} \log p(x_i, y_i|m, \hat{\theta}(\mathcal{D}_{-i})) \tag{7}$$

and $\hat{\theta}(\mathcal{D}_{-i})$ denotes the MLE computed on $\mathcal{D}$ excluding row i. Which model will CV pick and why? Hint: notice how the table of counts changes when you omit each training case one at a time.

e. Recall that an alternative to CV is to use the BIC score, defined as

$$BIC(M, \mathcal{D}) \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{MLE}) - \frac{\text{dof}(M)}{2} \log N \tag{8}$$

where $\text{dof}(M)$ is the number of free parameters in the model, Compute the BIC scores for both models (use log base e). Which model does BIC prefer?

Exercise 5.9 Posterior median is optimal estimate under L1 loss

Prove that the posterior median is optimal estimate under L1 loss.

Q6: Using an imperfect oracle

Consider a binary classification problem of predicting binary class y from features x. The cost of wrong prediction is \$6 and the cost of correct prediction is 0. Suppose the cost of asking a human to perform the manual classification is \$2. Therefore for a particular x, there are three possible decisions: (1) decision α_0 predicts y to be 0, (2) decision α_1 predicts y to be 1 and (3) decision α_h requires a human to perform the manual classification. Let $p_1 = p(y = 1|x)$

(i) Assume the human is 100% accurate and suppose $p_1 = 0.4$, what should our decision be to minimize expected loss?

(ii) Assume the human is 100% accurate and suppose $p_1 = 0.1$, what should our decision be to minimize expected loss?

(iii) Assume the human is 100% accurate. What is the general decision rule (as a function of p_1) in order to minimize expected loss?

(iv) Assume the human is only 95% accurate. What is the general decision rule (as a function of p_1) in order to minimize expected loss?