DD2447 Statistical Methods in Applied Computer Science, Fall 2020
Assignment 1 - Version 1.1

Jens Lagergren, Niharika Gauraha, Hazal Koptagel and Oskar Kviman

Deadline: see Canvas

You will present the assignment by a written report, submitted before the deadline using Canvas. You must solve the assignment individually and it will automatically be checked for similarities to other students' solutions as well as documents on the web in general. From the report it should be clear what you have done and you need to support your claims with results. You are supposed to write down the answers to the specific questions detailed for each task. This report should clearly show how you have drawn your conclusions and explain your derivations. Your assumptions, if any, should be stated clearly.

Being able to communicate results and conclusions is a key aspect of any scientific practitioner. It is up to you as a author to make sure that the report clearly shows what you have done. Based on this, and only this, we will decide if you pass the task. No detective work should be required on our side. In particular, neat and tidy reports please!

The grading of the assignment will be as follows,

E Correctly completed **one** task (out of 1.1-1.5)

D Correctly completed **two** tasks.

C Correctly completed **three** tasks.

B Correctly completed **four** tasks.

A Correctly completed **five** tasks.

These grades are valid for assignments submitted before the deadline, late assignments can at most receive the grade E. We allow us, whenever we find it appropriate, to also consider slightly incorrect solutions as correct.

Good Luck!

Version Log

- Version 1.1 - November 17, 2020. Typo fix in Question 5-a.

1.1 Graphical Models

Question 1: *Solve 2.7 from Murphy.*
Exercise 2.7 *Pairwise independence does not imply mutual independence*
We say that two random variables are pairwise independent if

$$p(X_2|X_1) = p(X_2)$$

and hence

$$p(X_2, X_1) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2)$$

We say that n random variables are mutually independent if

$$p(X_i|X_S) = p(X_i) \qquad \forall S \subseteq \{1, \dots, n\} \backslash \{i\}$$

and hence

$$p(X_{1:n}) = \prod_{i=1}^{n} p(X_i)$$

Show that pairwise independence between all pairs of variables does not necessarily imply mutual independence. It suffices to give a counter example.

Question 2: *Solve 2.8 from Murphy*
Exercise 2.8 *Conditional independence iff joint factorizes*
In the text we said $X \perp Y|Z$ iff

$$p(x, y|z) = p(x|z)p(y|z)$$

for all x, y, z such that $p(z) > 0$. Now prove the following alternative definition: $X \perp Y|Z$ iff there exist function g and h such that

$$p(x, y|z) = g(x, z)h(y, z)$$

for all x, y, z such that $p(z) > 0$.

1.2 Generative Models

Question 3: *Solve 3.11 from Murphy*
Exercise 3.11 *Bayesian analysis of the exponential distribution*
A lifetime X of a machine is modeled by an exponential distribution with unknown parameter θ. The likelihood is $p(x|\theta) = \theta e^{-\theta x}$ for $x \geq 0, \theta > 0$.

(a) Show that the MLE is $\hat{\theta} = 1/\bar{x}$, where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$.

(b) Suppose we observe $X_1 = 5, X_2 = 6, X_3 = 4$ (the lifetimes (in years) of 3 different iid machines). What is the MLE given this data?

(c) Assume that an expert believes θ should have a prior distribution that is also exponential $p(\theta) = Expon(\theta|\lambda)$. Choose the prior parameter, call it $\hat{\lambda}$, such that $\mathbb{E}[\theta] = 1/3$. Hint: recall that the Gamma distribution has the form $Ga(\theta|a, b) \propto \theta^{a-1}e^{-\theta b}$ and its mean is $\frac{a}{b}$.

(d) What is the posterior, $p(\theta|\mathcal{D}, \hat{\lambda})$?

(e) Is the exponential prior conjugate to the exponential likelihood?

(f) *What is the posterior mean,* $\mathbb{E}[\theta|\mathcal{D},\hat{\lambda}]$ *?*

(g) *Explain why the MLE and posterior mean differ. Which is more reasonable in this example?*

Question 4: *Solve 3.13 from Murphy*
Exercise 3.13 *Posterior predictive distribution for a batch of data with the dirichlet-multinomial model*
In Equation 3.51, we gave the the posterior predictive distribution for a single multinomial trial using a dirichlet prior $(p(X=j|D)=(\alpha_j+N_j)/(\alpha_0+N))$*. Now consider predicting a batch of new data,* $\widetilde{D}=(X_1,...,X_m)$*, consisting of m single multinomial trials (think of predicting the next m words in a sentence, assuming they are drawn iid). Derive an expression for*

$$p(\widetilde{D}|D,\boldsymbol{\alpha})$$

Your answer should be a function of $\boldsymbol{\alpha}$*, and the old and new counts (sufficient statistics), defined as*

$$N_k^{old}=\sum_{i\in D}I(x_i=k)$$

$$N_k^{new}=\sum_{i\in\widetilde{D}}I(x_i=k)$$

Hint: recall that, for a vector of counts, $N_{1:K}$*, the marginal likelihood (evidence) is given by*

$$p(D|\boldsymbol{\alpha})=\frac{\Gamma(\alpha)}{\Gamma(N+\alpha)}\prod_k\frac{\Gamma(N_k+\alpha_k)}{\Gamma(\alpha_k)}$$

where $\alpha=\sum_k\alpha_k$ *and* $N=\sum_k N_k$*.*

1.3 Bayesian Inference

Question 5: *Bayesian Inference for the Univariate Normal.*
In the standard form, the likelihood has two parameters, the mean μ *and the variance* σ^2*:*

$$p(x_1,x_2,\ldots,x_n|\mu,\sigma^2)\propto\frac{1}{\sigma^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2\right)$$

It is desired to find conjugate prior distributions for these parameters.

(a) *Assuming* σ^2 *is fixed, show that the conjugate prior for* μ *is a Gaussian distribution. If we assume:*

$$x_i|\mu,\tau\sim\mathcal{N}(\mu,\sigma^2),iid$$
$$\mu|\mu_0,\sigma_0^2\sim\mathcal{N}(\mu_0,\sigma_0^2)$$

Then the posterior is:

$$\mu|x\sim\mathcal{N}\left(\frac{\sigma_0^2n}{\sigma^2+n\sigma_0^2}\bar{x}+\frac{\sigma^2}{\sigma^2+n\sigma_0^2}\mu_0,\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}\right)$$

(b) *Keeping μ fixed, show that the conjugate prior for σ^2 is an inverse Gamma distribution. If we assume:*

$$
\begin{aligned}
x_i|\mu, \sigma^2 &\sim \mathcal{N}(\mu, \sigma^2), iid \\
\sigma^2|\alpha, \beta &\sim \mathcal{IG}(\alpha, \beta)
\end{aligned}
$$

Then the posterior is:

$$\sigma^2|x \sim \mathcal{IG}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right)$$

(c) *Both the mean μ and variance σ^2 are unknown. If we assume:*

$$
\begin{aligned}
x_i|\mu, \tau &\sim \mathcal{N}(\mu, \tau^{-1}), iid \\
\mu|\tau &\sim \mathcal{N}(\mu_0, (n_0\tau)^{-1}) \\
\tau &\sim Ga(\alpha, \beta)
\end{aligned}
$$

where $\tau = 1/\sigma^2$ is the precision term. Then the posterior is:

$$
\begin{aligned}
\mu|\tau, x &\sim \mathcal{N}\left(\frac{n}{n+n_0}\bar{x} + \frac{n_0}{n+n_0}\mu_0, (n\tau + n_0\tau)^{-1}\right) \\
\tau|x &\sim Ga\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{nn_0}{2(n+n_0)}(\bar{x} - \mu_0)^2\right)
\end{aligned}
$$

Question 6: *Solve 3.18 from Murphy*
Exercise 3.18 *Bayes factor for coin tossing*
Suppose we toss a coin $N = 10$ times and observe $N_1 = 9$ heads. Let the null hypothesis be that the coin is fair, and the alternative be that the coin can have any bias, so $p(\theta) = Unif(0, 1)$. Derive the Bayes factor $BF_{1,0}$ in favor of the biased coin hypothesis. What if $N = 100$ and $N_1 = 90$? Hint: see Exercise 3.17.

1.4 Mixture and MLE

Question 7: *Solve 5.1 from Murphy*
Exercise 5.1 *Proof that a mixture of conjugate priors is indeed conjugate*
Derive Equation 5.69 ($p(\theta|D) = \sum_k p(z = k|D)p(\theta|D, z = k)$).

Question 8: *Solve 5.8 from Murphy*
Exercise 5.8 *MLE and model selection for a 2d discrete distribution*
(Source: Jaakkola.)
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: $x = (1, 1, 0, 1, 1, 0, 0)$, $y = (1, 0, 0, 0, 1, 0, 1)$. What are the MLEs for θ_1 and θ_2? Justify your answer. Hint: note that the likelihood function factorizes,*

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

What is $p(D|\hat{\boldsymbol{\theta}}, M_2)$ 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 $p(x, y|\boldsymbol{\theta}) = \theta_{x,y}$. (Only 3 of these parameters are free to vary, since they must sum to one.) What is the MLE of $\boldsymbol{\theta}$? What is $p(D|\hat{\boldsymbol{\theta}}, M_4)$ 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:*

$$L(m) = \sum_{i=1}^{n} \log p(x_i, y_i|m, \hat{\theta}(D_{-i}))$$

and $\hat{\theta}(D_{-i})$ denotes the MLE computed on 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*

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

where $\mathrm{dof}(M)$ 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?

1.5 Numerical Problems

Question 9: *Solve 18 from Grinstead*
Chapter 4 Exercise 18
A doctor assumes that a patient has one of three diseases d_1, d_2, or d_3. Before any test, he assumes an equal probability for each disease. He carries out a test that will be positive with probability 0.8 if the patient has d_1, 0.6 if he has disease d_2, and 0.4 if he has disease d_3. Given that the outcome of the test was positive, what probabilities should the doctor now assign to the three possible diseases?

Question 10: *Solve 3.14 from Murphy*
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* $p(x_{2001} = e|D)$*, if we assume* $\boldsymbol{\theta} \sim Dir(\alpha_1, \dots, \alpha_{27})$*, 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* $p(x_{2001} = p, x_{2002} = a|D)$*, if we assume* $\boldsymbol{\theta} \sim Dir(\alpha_1, \dots, \alpha_{27})$*, where* $\alpha_k = 10$ *for all* k*? Show your work.*