DS 102 Homework 3

If you are handwriting your solution please make sure your answers are legible, as you may lose points otherwise.
Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you do discuss the homework with others, please include their names on your submission.
Due on Gradescope by 9:29am, Thursday March 5th, 2020

1. (20 points) Here we'll work through an example of maximum *a posteriori* (MAP) estimation, and look at one interpretation of MAP linear regression weights.

   Suppose $x_1, \ldots, x_n \in \mathbb{R}^d$ are fixed feature vectors. We assume the linear Gaussian model, where we observe

   $$y_i = \beta^\top x_i + \epsilon_i,\ i = 1, \ldots, n,$$

   where $\epsilon_i \sim N(0, \sigma^2)$ are independent of each other, and $\beta \in \mathbb{R}^d$ and $\sigma^2 > 0$ are unknown.

   Let $y = (y_1, \ldots, y_n), \epsilon = (\epsilon_1, \ldots, \epsilon_n)$, and let X denote the matrix whose i-th row is equal to x_i. Using this notation, we have

   $$y = X\beta + \epsilon.$$

   We model the regression weights as a random variable with the following prior distribution:

   $$\beta \sim N(0, \sigma_\beta^2 \cdot I).$$

   where $\sigma_\beta^2 > 0$ is hyperparameter we choose. That is, every entry of β is distributed as $N(0, \sigma_\beta^2)$ and the entries are independent.

   (a) (8 points) Write the posterior distribution for β after observing the data, $p(\beta|X, y)$. It's fine to just derive an expression that the posterior is proportional to. (Hints: use Bayes' rule and the probability density functions of multivariate Gaussians. Also use the fact that for a vector z, $z^T z = \|z\|_2^2$, where $\|z\|_2$ is the Euclidean norm of z.)

   (b) (8 points) Show that the MAP estimator of β,

   $$\hat{\beta}_{MAP} := \arg\max_\beta p(\beta|X, y)$$

   is also the solution to the regularized least-squares problem,

   $$\arg\min_\beta \|X\beta - y\|_2^2 + \lambda\|\beta\|_2^2$$

   for $\lambda = \frac{\sigma^2}{\sigma_\beta^2}$. (Hints: use the expression for the posterior you derived in the previous part, and the fact that taking the logarithm of a function does not change its argmax. That is, $\arg\max_\beta f(\beta) = \arg\max_\beta \log f(\beta)$. Multiplying a function by a constant also does not change its argmax.)

(c) (4 points) In the regularized least-squares problem, λ is the regularization term: large values of λ penalize weight vectors with large norms. Since $\hat{\beta}_{MAP}$ is the solution to the regularized least-squares problem with $\lambda = \frac{\sigma^2}{\sigma_\beta^2}$, explain how our modeling decisions (*i.e.*, our choice of σ_β^2) influences our solution $\hat{\beta}_{MAP}$.

2. (30 points) **EM for Poisson Mixture Model**

Suppose CalTrans wants to study the probability of traffic accidents in the Bay Area so it can improve highway infrastructure. They collect a dataset $X = (x_1, ..., x_n)$ of the number of accidents that occur in a day, for n days.

We want to model each count as being drawn from a Poisson distribution. However, we believe the mean of that Poisson depends on whether that day is sunny, raining, or snowing, which happens with some probabilities π_1, π_2, π_3 such that $\sum_j \pi_j = 1$, respectively. That is, we model each count as follows:

$$z_i \sim \text{Multinomial}(1, (\pi_1, \pi_2, \pi_3))$$
$$x_i \sim \text{Poisson}(\lambda_{z_i})$$

where z_i is a categorical variable that determines whether it is is sunny, raining, or snowing, and $\lambda_1, \lambda_2, \lambda_3$ are the means of the Poisson distributions when it is sunny, raining, and snowing, respectively. Assume the z_i, and therefore the counts x_i, are independent of each other.

Unfortunately, CalTrans did not record the weather z_i corresponding to each count, so z_i is an unobserved latent variable. In this problem, we'll work out how to estimate $z = (z_1, \dots z_n)$, $\pi = (\pi_1, \pi_2, \pi_3)$, and $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ using expectation-maximization (EM).

(a) (3 points) Suppose we did know z_i for count x_i. Given the parameters λ, derive the likelihood of observing x_i conditioned on z_i, $\mathbb{P}(x_i | z_i = j, \lambda)$.

(b) (2 points) Given the parameters π and λ, derive the joint likelihood $\mathbb{P}(x_i, z_i = j \mid \pi, \lambda)$. Use your result from the previous part.

(c) (5 points) Before we dive into EM, first write down the likelihood $\mathbb{P}(x \mid \lambda, \pi)$. (Hint: use the assumption that the x_i are independent from each other, and your result from the previous part.)

(d) (10 points) The expression you derived in the previous part has no closed-form solution for the maximum likelihood estimates (MLEs) of λ, π. We now turn to EM to approximate the MLEs of λ, π.

(i) For the E-step at iteration t, we have our current estimates $\lambda_1^{(t)}, \lambda_2^{(t)}, \lambda_3^{(t)}$ and $\pi_1^{(t)}, \pi_2^{(t)}, \pi_3^{(t)}$. Derive the posterior distribution of z_i conditioned on x_i and these current estimates, $\mathbb{P}(z_i = j \mid x_i, \lambda^{(t)}, \pi^{(t)})$. We'll denote this posterior distribution $q_i^{(t)}(z_i)$, which is a multinomial distribution since z_i is a categorical random variable. (Hint: Use Bayes' theorem.)

(ii) Your result $q_i^{(t)}(z_i)$ is the posterior over a single z_i given x_i. What is the posterior over all $z = (z_1, \ldots, z_n)$ given all $x = (x_1, \ldots, x_n)$, $\mathbb{P}(z \mid x, \lambda^{(t)}, \pi^{(t)})$? This is the $q^{(t)}(z)$ that results from the E-step. (You can write it in terms of $q_i^{(t)}(z_i)$.)

(e) (10 points) To derive the M-step, using the independence assumptions we have

$$\pi^{(t+1)}, \lambda^{(t+1)} = \underset{\lambda,\pi:\, \sum_j \pi_j = 1}{\arg\max}\ \mathbb{E}_{z\sim q^{(t)}}[\log \mathbb{P}(x, z \mid \lambda, \pi)] \tag{1}$$

$$= \underset{\lambda,\pi:\, \sum_j \pi_j = 1}{\arg\max}\ \mathbb{E}_{z\sim q^{(t)}}[\log \prod_{i=1}^{n} \mathbb{P}(x_i, z_i \mid \lambda, \pi)] \tag{2}$$

$$= \underset{\lambda,\pi:\, \sum_j \pi_j = 1}{\arg\max}\ \mathbb{E}_{z\sim q^{(t)}}[\sum_{i=1}^{n} \log \mathbb{P}(x_i, z_i \mid \lambda, \pi)] \tag{3}$$

$$= \underset{\lambda,\pi:\, \sum_j \pi_j = 1}{\arg\max} \sum_{i=1}^{n} \mathbb{E}_{z\sim q^{(t)}}[\log \mathbb{P}(x_i, z_i \mid \lambda, \pi)] \tag{4}$$

$$= \underset{\lambda,\pi:\, \sum_j \pi_j = 1}{\arg\max} \sum_{i=1}^{n} \mathbb{E}_{z_i\sim q_i^{(t)}}[\log \mathbb{P}(x_i, z_i \mid \lambda, \pi)] \tag{5}$$

$$= \underset{\lambda,\pi:\, \sum_j \pi_j = 1}{\arg\max} \sum_{i=1}^{n} \mathbb{E}_{z_i\sim q_i^{(t)}}[\log \mathbb{P}(x_i \mid z_i)\mathbb{P}(z_i \mid \pi)] \tag{6}$$

$$= \underset{\lambda,\pi:\, \sum_j \pi_j = 1}{\arg\max} \sum_{i=1}^{n} \mathbb{E}_{z_i\sim q_i^{(t)}}[\log \left(\pi_{z_i} \frac{\lambda_{z_i}^{x_i}}{x_i!} \exp(-\lambda_{z_i})\right)]. \tag{7}$$

It's a good check to make sure you understand this derivation. Importantly, note that due to independence, we just need to consider the expectation with respect to the posterior $q_i^{(t)}$ over one z_i (which you derived in Part d(i)) instead of over all z at once. Also note that in the maximization, we need to satisfy the constraint that $\sum_{j=1}^{3} \pi_j = 1$ in order to have valid probabilities.

(i) We can solve for $\pi^{(t+1)}, \lambda^{(t+1)}$ by taking the partial derivatives of

$$f(\pi, \lambda) = \mathbb{E}_{z\sim q^{(t)}}[\log \mathbb{P}(x, z \mid \lambda, \pi)] = \sum_{i=1}^{n} \mathbb{E}_{z_i\sim q_i^{(t)}}[\log \left(\pi_{z_i} \frac{\lambda_{z_i}^{x_i}}{x_i!} \exp(-\lambda_{z_i})\right)]$$

with respect to components λ_j and π_j, and setting them to zero.

Derive the partial derivative $\frac{\partial f(\pi,\lambda)}{\partial \lambda_j}$ and set it to zero. What is $\lambda_j^{(t+1)}$? (Hints: You can pass the derivative through both the sum and the expectation. You can also use the fact that $\mathbb{E}_{z_i\sim q_i^{(t)}}[\mathbb{1}(z_i = j)] = q_i^{(t)}(z_i = j)$.)

(ii) Since we need to satisfy the constraint that $\sum_{j=1}^{3} \pi_j = 1$, we can't just set $\frac{\partial f(\pi,\lambda)}{\partial \pi_j}$ to zero to get $\pi_j^{(t+1)}$. Instead, we need to use Lagrange multipliers to solve the constrained optimization problem. This turns out to be equivalent to setting $\frac{\partial f(\pi,\lambda)}{\partial \pi_j} - n = 0$.

Derive the partial derivative $\frac{\partial f(\pi,\lambda)}{\partial \pi_j}$. Using $\frac{\partial f(\pi,\lambda)}{\partial \pi_j} - n = 0$, what is $\pi_j^{(t+1)}$?