CSE 291: Unsupervised learning

Homework 6

This homework is due on Thursday May 25 at 11.59pm.

- All homeworks must be typewritten and uploaded to Gradescope.
- No late homeworks will be accepted.

1. *Poisson and exponential.* The exponential(λ) distribution, introduced earlier, is commonly used to model waiting times: the elapsed time before the phone next rings, the elapsed time before disaster strikes again, and so on.

   (a) Show that the exponential distribution has a *memoryless* property: If $X \sim$ exponential(λ), then for $s, t \geq 0$,
   $$\Pr(X > s + t | X > s) \;=\; \Pr(X > t).$$
   This means, for instance, that if you're waiting for the phone to ring, the probability of having to wait another hour does not depend on how long you have already waited.

   (b) Suppose the waiting time for a particular event follows an exponential distribution with mean 1. For any positive integer k, let Z_k denote the waiting time for k successive events. What is the distribution of Z_k? You should use the following useful fact: if $X \sim$ gamma(α_1, β) and $Y \sim$ gamma(α_2, β) are independent, then their sum, $X+Y$, has a gamma($\alpha_1+\alpha_2, \beta$) distribution. *Hint:* We related the exponential to the gamma in an earlier homework.

   (c) Continuing from part (b), show that the number of events that occur in a particular amount of time t follows a Poisson distribution. What is the mean of this Poisson, as a function of t?

2. *Change of variable and random generation.* Suppose random variable $X \in \mathbb{R}$ has distribution p. What is the distribution of $Y = f(X)$, where $f : \mathbb{R} \to \mathbb{R}$ is a one-to-one (and hence invertible) function? If X is discrete, then the distribution q of Y is simply
   $$q(y) = p(g(y))$$
   where $g : \mathbb{R} \to \mathbb{R}$ is the inverse of f. If X is continuous, then a scaling term is also needed:
   $$q(y) = |g'(y)|\, p(g(y)).$$
   Using these results, characterize the distribution of the variable Y generated as follows:

   - Pick U at random from the uniform distribution over $[0, 1]$.
   - Set $Y = -(\ln U)/\beta$, where $\beta > 0$ is some fixed parameter.

3. In an election, there are three candidates, A, B, and C. Of the voting population, θ_A fraction will vote for A, while θ_B fraction will vote for B and θ_C fraction will vote for C; thus $\theta_A + \theta_B + \theta_C = 1$.

(a) John wishes to infer the distribution $(\theta_A, \theta_B, \theta_C)$. He starts by putting a uniform prior on it; that is, he considers all distributions to be equally likely. He then picks four people at random to poll, and finds that they intend to vote for A, B, B, A, respectively. Given this information, what is John's posterior distribution?

(b) Under this posterior distribution, what is the expected value of $(\theta_A, \theta_B, \theta_C)$?

4. *Rejection sampling for a bivariate distribution.* Suppose you are given a black-box function U that returns samples from the uniform distribution over $[0, 1]$. Design a rejection-sampling scheme that generates random samples from the distribution over $[0, 1]^2$ with density

$$p(x) \propto \exp(-\|x\|_1).$$

What is the expected number of calls to U in order to generate one sample from p?

5. Suppose we want to generate uniform-random points in the d-dimensional unit ball $B = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$. Here's a rejection-sampling approach:

- Pick a uniform-random point $X \in [-1, 1]^d$ by choosing each coordinate $X_i \in [-1, 1]$ independently.
- If $\|X\| \leq 1$: halt and output X; otherwise repeat.

In expectation, how many points will we need to pick in $[-1, 1]^d$ before we are able to output a point in B? Is this a good way of sampling from the unit ball?

6. *Modeling temperature data with Gaussian processes.* In this question, we will explore modeling of geospatial data via Gaussian processes. Begin by downloading the files `gptrain.csv` and `gptest.csv` from the course webpage. Each row in each file corresponds to a temperature reading at a given weather station in Brazil at 2pm on January 1, 2021. The first column gives the `latitude` of the station, the second the `longitude`, and the final column is the `temperature` reading in degrees Celsius. We will fit a simple Gaussian Process model on `gptrain.csv` and use it to infer temperatures at other locations at that same date and time.

We will fit a Gaussian process with covariance function

$$k(x, x') = \exp\left(\frac{-\|x - x'\|}{\sigma}\right),$$

and mean function $m(x) = 20$ (degrees Celsius). Here $\|x - x'\|$ denotes Euclidean distance between raw latitude and longitude coordinates.

Denote the locations in `gptrain.csv` by X_{tr} and the temperatures at those locations by y_{tr}; define X_{te}, y_{te} analogously for `gptest.csv`. Thus X_{tr} is a 93×2 matrix while y_{tr} is a 93-dimensional vector.

(a) The joint distribution of training and test responses can be written as

$$\begin{pmatrix} y_{\text{tr}} \\ y_{\text{te}} \end{pmatrix} \sim N\left(\begin{bmatrix} 20 \\ \vdots \\ 20 \end{bmatrix}, \begin{bmatrix} K_{\text{tr}} & K_{\text{tr,te}} \\ K_{\text{te,tr}} & K_{\text{te}} \end{bmatrix}\right),$$

where the block matrix

$$\begin{bmatrix} K_{\text{tr}} & K_{\text{tr,te}} \\ K_{\text{te,tr}} & K_{\text{te}} \end{bmatrix}$$

is the matrix arising from all pairwise evaluations of the covariance function $k(\cdot)$ over all training and testing locations, and its diagonal components K_{tr} and K_{te} are the matrices arising from all pairwise covariances within the training and test sets, respectively.

Explain how the definition of Gaussian process allows us to come to this conclusion.

(b) Define

$$m_{\text{te}} := \begin{bmatrix} 20 \\ \vdots \\ 20 \end{bmatrix} \in \mathbb{R}^{11},$$

since there are 11 locations in the test set. Using properties of the Gaussian, it can be shown that the distribution of test responses given training locations, training responses, and test locations, follows

$$\begin{aligned} & y_{\text{te}} | y_{\text{tr}}, X_{\text{te}}, X_{\text{tr}} \\ & \sim m_{\text{te}} + N\left(K_{\text{te,tr}} \cdot K_{\text{tr}}^{-1}\left(y_{\text{tr}} - m_{\text{te}}\right), \quad K_{\text{te}} - K_{\text{te,tr}} K_{\text{tr}}^{-1} K_{\text{tr,te}}\right), \end{aligned}$$

In Bayesian language, this is the posterior predictive distribution.

- Using X_{tr}, X_{te}, and y_{tr}, compute the posterior mean of y_{te} via matrix multiplication, using $\sigma = 1.5$ in the covariance function. Report the mean squared error between your computed posterior mean and the true values found in y_{te}.
- Report the mean squared error that you would have obtained by simply predicting the average value in y_{tr} for all locations.

(c) The diagonal of the posterior predictive matrix

$$K_{\text{te}} - K_{\text{te,tr}} K_{\text{tr}}^{-1} K_{\text{tr,te}}$$

gives the conditional variances of the test responses given X_{tr}, X_{te}, and y_{tr}.

- Create a rich grid (at least 5000 points) of test latitudes and longitudes within the range found in the training set. Make a plot showing the predicted (posterior mean) temperature at *all points in the grid.* You might find the function `matplotlib.pyplot.imshow` helpful for this.
- Make another plot showing the standard deviation of the prediction at *all points in the grid.* Plot the locations of the stations found in the training set as well.
- What do you notice about the relationship of the standard deviation of the prediction and the distance of that prediction to a training point?