ASSIGNMENT 2

Mathematics for Machine Learning 811 — 24 January 2022

Problems 1, 2, 4(a), 5(a) and 5(b) are intended for pen-and-paper, and your submission should include steps. The rest of the problems can be done with the aid of Python.

What you submit must be your own work, and sources other than the lecture material must be cited. Remember to append a signed plagiarism declaration to your submission.

---

1. Consider the bivariate distribution $p(x, y)$ of two discrete random variables X and Y, given in the table on the right.

	y_1	0.01	0.02	0.03	0.10	0.10
Y	y_2	0.05	0.10	0.05	0.07	0.20
	y_3	0.10	0.05	0.03	0.05	0.04
		x_1	x_2	x_3	x_4	x_5
				X		

(a) Find the marginal distributions $p(x)$ and $p(y)$.

(b) Find the conditional distribution $p(x \,|\, Y = y_1)$.

(c) Suppose the value of x_i is i, and the value of y_i is i. Compute the correlation between X and Y.

2. In a factory there are three machines that make light bulbs. The machines manufacture 20%, 30% and 50% of the total production. From their production, 5%, 4%, and 2% respectively are faulty. I choose a collection of light bulbs at random from the factory's output.

(a) If the collection contains two faulty light bulbs, what is the probability that those two come from the same machine?

(b) If the collection contains three faulty light bulbs, what is the probability that those three come from three different machines?

3. Generate 1,000 samples from the Gaussian distribution $\mathcal{N}\left(\begin{bmatrix}3\\2\end{bmatrix}, \begin{bmatrix}10 & 5\\5 & 5\end{bmatrix}\right)$ by means of

(a) the procedure given in section 6.5.4 of the textbook;

(b) the probability integral transform (see paragraph immediately after Theorem 6.15 in the book).
Hint: the inverse Gaussian cdf can be written in terms of the inverse error function (`scipy.special.erfinv`).

For each method separately, plot the samples, calculate the empirical mean and covariance, and compare with the true mean and covariance.

4. Each of the N rows in `x.dat` is a training sample $\boldsymbol{x}_n \in \mathbb{R}^2$ with corresponding label $y_n \in \{0, 1\}$ in `y.dat`. Your task will be to use gradient descent in order to fit a logistic classifier,

$$\sigma(\boldsymbol{x}) = \frac{1}{1 + \exp(-\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)},$$

to this data. The model parameters are $\boldsymbol{w} \in \mathbb{R}^2$ and $b \in \mathbb{R}$, and a sensible choice for the objective function in this case is the "cross-entropy loss" between true labels y_n and predicted labels $\sigma(\boldsymbol{x}_n)$, averaged over the training set:

$$L(\boldsymbol{w}, b) = \frac{1}{N}\sum_{n=1}^{N}\Big[-y_n \log\big(\sigma(\boldsymbol{x}_n)\big) - (1 - y_n)\log\big(1 - \sigma(\boldsymbol{x}_n)\big)\Big].$$

p.t.o.

(a) Prove that the gradient descent update rules for minimising $L(\boldsymbol{w}, b)$ are as follows:

$$\boldsymbol{w}_{i+1} = \boldsymbol{w}_i - \gamma \frac{1}{N} \sum_{n=1}^{N} \big(\sigma(\boldsymbol{x}_n) - y_n\big)\boldsymbol{x}_n, \quad b_{i+1} = b_i - \gamma \frac{1}{N} \sum_{n=1}^{N} \big(\sigma(\boldsymbol{x}_n) - y_n\big).$$

(b) Initialise $\boldsymbol{w}_0$ and b_0 with zeros, and implement batch gradient descent. A fixed value for γ between 1 and 5 seems reasonable, but you are free to experiment. To check convergence, plot $L(\boldsymbol{w}_i, b_i)$ as a function of i. Give your final values of $\boldsymbol{w}$ and b.

(c) Plot all $\boldsymbol{x}_n$ points in the x_1–x_2 plane, using two different colours for $y_n = 0$ and $y_n = 1$, as well as the decision boundary of the trained logistic model. Here the decision boundary would consist of all $\boldsymbol{x}$ for which $\sigma(\boldsymbol{x}) = 0.5$, that is, $\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b = 0$ (a straight line in the x_1–x_2 plane that tries to separate the two classes).

5. Consider a coin for which the probability μ of landing on heads is unknown, and let X be a binomial random variable representing the number of heads in N flips of this coin.

(a) Determine the maximum likelihood estimate of μ, given that h of the N flips resulted in heads.

Hint: differentiate $p(x = h \,|\, N, \mu)$ with respect to μ, set it to 0, and solve for θ.

(b) Place a beta prior on μ, with parameters α and β. See Example 6.11 on page 208 of the textbook. Write down the complete posterior distribution $p(\mu \,|\, x = h, N, \alpha, \beta)$, specifying all constants omitted by the book (necessary for the plots in part (c)). Then determine the maximum a posteriori estimate of μ.

(c) Choose some value for μ in $(0, 1)$, simulate 100 flips of the coin, and let h be the number of heads obtained. Now again suppose μ is unknown. Plot the prior distribution $p(\mu \,|\, \alpha, \beta)$ and posterior distribution $p(\mu \,|\, x = h, N, \alpha, \beta)$ from part (b) as functions of μ, for each of the following parameter choices:

i. $\alpha = 1$, $\beta = 1$ (a uniform prior)

ii. $\alpha = 8$, $\beta = 8$ (a unimodal prior centred around $\mu = \frac{1}{2}$).

How do the graphs compare with the true value of μ?

6. Your task here will be to generate a noisy dataset and then apply linear regression to fit polynomials of various degrees, similar to what is shown in Figures 9.4, 9.5 and 9.7 of the textbook.

(a) Generate $N = 10$ data points (x_n, y_n), with x_n sampled uniformly randomly from $[-5, 5]$, and $y_n = \sin(x_n/5) + \cos(x_n) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.04)$. Plot your data as dots in the x–y plane. This will be the fixed dataset for parts (b) and (c) below.

(b) Fit polynomials of degree $1, 2, \ldots, 9$ using maximum likelihood estimation. Plot each one separately over the data points (similar to Figure 9.5).

Note: when fitting a polynomial of degree p, create a data vector $\boldsymbol{x}_n = [1,\ x_n,\ x_n^2,\ \ldots,\ x_n^p]^T$ for each input x_n.

(c) Repeat part (b) using maximum a posteriori estimation, with Gaussian priors $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{0}, b^2\boldsymbol{I})$. Briefly discuss and compare to your graphs in part (b).