NYU Paris **Spring 2019**

CSCI-UA 9473 Introduction to Machine Learning

Problem Set 1: Probability, Statistics and Inference

Exercise 1 (source: Minka)**.** *My neighbor has two children. Assuming that the gender of a child is like a coin flip, it is most likely, a priori, that my neighbor has one boy and one girl, with probability* $1/2$*. The other possibilities-two boys or two girls-both have probabilities* $1/4$*.*

- *Suppose I ask him whether he has any boys, and he says yes. What is the probability that one child is a girl?*
- *Now suppose that instead of asking him, I happen to see one of his children run by, and it happens to be a boy. What is the probability that the other child is a girl?*

Exercise 2. *Show that* $\mathbb{E}\{var(u|v)\} + var(\mathbb{E}\{u|v\}) = var(u)$

Exercise 3 (source: Murphy)**.** *The normalization constant for a zero mean Gaussian is given by*

$$Z = \int_a^b \exp(-\frac{x^2}{2\sigma^2})\, dx \tag{1}$$

where $a = -\infty$ *and* $b = \infty$*. To compute this, consider its square*

$$Z^2 = \int_a^b \int_a^b \exp(-\frac{x^2+y^2}{2\sigma^2})\, dx\, dy \tag{2}$$

Then, change variables from cartesian (x, y) *to polar* (r, θ) *using* $r = \cos\theta$ *and* $y = r\sin\theta$*. Since* $dx\, dy = r\, dr\, d\theta$*, and* $\cos^2\theta + \sin^2\theta = 1$*, we have*

$$Z^2 = \int_0^{2\pi} \int_0^{\infty} r\exp(-\frac{r^2}{2\sigma^2}) dr\, d\theta \tag{3}$$

Evaluate the integral and hence show $Z = \sqrt{\sigma^2 2\pi}$*. Hint: consider the derivative of the integrand.*

Exercise 4 (source: Murphy)**.** *We say that two random variables are pairwise independent if*

$$p(X_2|X_1) = p(X_2) \tag{4}$$

and hence

$$p(X_2, X_1) = p(X_1)p(X_2|X_1) = p(X_1)p(X_2) \tag{5}$$

We say that n *random variables are mutually independent if*

$$p(X_i|X_S) = p(X_i), \quad \forall S \subseteq \{1, \ldots, n\} \setminus \{i\} \tag{6}$$

and hence

$$p(X_{1:n}) = \prod_{i=1}^{n} p(X_i) \tag{7}$$

Show that pairwise independence between all pairs of variables does not necessarily imply mutual independence. Hint: consider the event space $\Omega = \{1, 2, 3, 4\}$ *where the probability of getting each of the four numbers is* $1/4$*. Consider the set of events* $A = \{1, 2\}$, $B = \{1, 3\}$ *and* $C = \{1, 4\}$*. What is the probability of each event? To show pairwise independence, consider the events* $A \cap B$, $B \cap C$ *and* $A \cap C$ *corresponding to the probability of getting an element which belongs to both subsets. Then consider the probability of the intersection* $A \cap B \cap C$

Exercise 5 (source: Koller)**.** *After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease and that the test is* 99% *accurate (i.e the probability of testing positive given that you have the disease is* .99 *as is the probability of testing negative given that you don't have the disease). The good news is that this is a rare disease striking only on in 10,000 people. What are the chances that you actually have the disease?*

Exercise 6 (source: CB)**.** *The exponential family has interesting properties that are useful in machine learning. Show that each of the following families is an exponential family*

- *normal family with either parameter* μ *or* σ *known*
- *Gamma family with either parameter* α *or* β *known or both unknown*
- *Beta family with either parameter* α *or* β *known or both unknown*
- *Poisson family*
- *negative binomial family with* r *known,* $0 < p < 1$*.*

Exercise 7 (source: CB)**.** *A man with* n *keys wants to open his door and tries the keys at random. Exactly one key will open the door. Find the mean number of trials if*

- *unsuccessful keys are not eliminated from further selections*
- *unsuccesful keys are eliminated*

Exercise 8 (source: CB)**.** *Let the number of chocolate chips in a certain type of cookie have a Poisson distribution. We want the probability that a randomly chosen cookie has at least two chocolate chips to be greater than .99. Find the smallest value of the mean of the distribution that ensures this probability.*

Exercise 9. *The Maximum likelihood estimator (MLE) which is often denoted as* $\hat{\theta}(\boldsymbol{x})$ *is the parameter value at which the likelihood function* $L(\theta|\boldsymbol{x})$ *attains its maximum as a function of* θ *with* $\boldsymbol{x}$ *fixed. If the likelihood function is differentiable, possible candidates for the MLE are the values of* $(\theta_1, \ldots, \theta_k)$ *that solve*

$$\frac{\partial}{\partial \theta_i} L(\theta|\boldsymbol{x}) = 0 \tag{8}$$

Let $X_1, \ldots, X_n$ *be i.i.d Bernoulli(p). Then the likelihood function is*

$$L(p|\boldsymbol{x}) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^y(1-p)^{n-y},$$

where $y = \sum x_i$*. Compute the MLE.*

Exercise 10. *Let $X_1, \ldots, X_n$ be i.i.d normal random variables with fixed variance. I.e, $X_i \sim \mathcal{N}(\theta, 1)$. Give the expression of the likelihood function. Then compute the MLE. Set the derivative to zero and prove that you have a maximum.*

Exercise 11. *To verify that a function of two variables $H(\theta_1, \theta_2)$ has a local maximum at $(\hat{\theta}_1, \hat{\theta}_2)$, it must be shown that the following three conditions hold:*

- *First order partial derivatives must vanish*

$$\left.\frac{\partial}{\partial\theta_1}H(\theta_1,\theta_2)\right|_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2} = 0, \quad \textit{and} \quad \left.\frac{\partial}{\partial\theta_2}H(\theta_1,\theta_2)\right|_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2} = 0 \tag{9}$$

- *At least one second order partial derivative is negative*

$$\left.\left|\frac{\partial^2}{\partial\theta_1^2}H(\theta_1,\theta_2)\right|\right|_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2} < 0, \quad \textit{or} \quad \left.\left|\frac{\partial^2}{\partial\theta_2}H(\theta_1,\theta_2)\right|\right|_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2} < 0 \tag{10}$$

- *The Jacobian of the second-order partial derivatives is positive*

$$\left.\begin{vmatrix} \frac{\partial^2}{\partial\theta_1^2}H(\theta_1,\theta_2) & \frac{\partial^2}{\partial\theta_1\partial\theta_2}H(\theta_1,\theta_2) \\ \frac{\partial^2}{\partial\theta_1\partial\theta_2}H(\theta_1,\theta_2) & \frac{\partial^2}{\partial\theta_2^2}H(\theta_1,\theta_2) \end{vmatrix}\right|_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2} \tag{11}$$

$$= \frac{\partial^2}{\partial\theta_1^2}H(\theta_1,\theta_2)\frac{\partial^2}{\partial\theta_2^2}H(\theta_1,\theta_2) - \left(\frac{\partial^2}{\partial\theta_1\partial\theta_2}H(\theta_1,\theta_2)\right)^2\Bigg|_{\theta_1=\hat{\theta}_1,\theta_2=\hat{\theta}_2} > 0 \tag{12}$$

Compute the normal MLE when both the mean and the variance are unknown. Then show that your estimator is a local maximum.

Exercise 12 (source: Jeffrey Miller, Harvard)**.** *We consider an uncalibrated sensor from which we acquire data $X \sim Uniform(0, \theta)$ for some unknown $\theta > 0$. I.e*

$$p(x|\theta) = \begin{cases} 1/\theta & \textit{if } 0 < x < \theta \\ 0 & \textit{otherwise} \end{cases} \tag{13}$$

$$= \frac{1}{\theta} I_{0,\theta}(x) \tag{14}$$

where $I_{(0,\theta)}(x)$ is the indicator function which equals 1 when $0 < x < \theta$ and 0 otherwise. We acquire samples $D = (x_1, \ldots, x_n)$. We would like to estimate the underlying parameter to characterize the distribution.

- *Write down the likelihood function for $p(D|\theta)$. Find the maximum likelihood estimator for θ. As above, prove that your estimator is indeed the MLE.*
- *We now consider the following prior distribution on θ.*

$$p(\theta) = \alpha\beta^\alpha\theta^{-\alpha-1}I_{\beta,\infty}(\theta) \tag{15}$$

  which is known as a Pareto distribution. Plot the prior densities corresponding to the parameters $(\alpha, \beta) = \{(0.1, 0.1), (2.0, 0.1), (1.0, 2.0)\}$
- *We still assume that the data follows a uniform distribution ($X_1, \ldots, X_n \sim Uniform(0, \theta)$) but we consider uncertainty on the parameter which we encode through the Pareto distribution, i.e. $\theta \sim Pareto(\alpha, \beta)$. Write down the posterior distribution $p(\theta|D)$. Does it belong to a family of distribution you know?*
- *Using the posterior from the previous point, derive the MAP for θ.*

- *The square loss is defined as* $L(\theta, \hat{\theta})$. *For the Pareto posterior, what estimator of* θ *minimizes the posterior expected loss (i.e. the posterior loss averaged over all possible values of* θ*)? Simplify your answer as much as possible. How does it compare to the MLE? What about the MAP?*
- *Suppose the data you observe is given by* $D = \{0.7, 1.3, 1.7\}$. *Plot the density of the posterior distribution of* θ *for each of the three Pareto priors used in the second point. For each of those three priors, what is the MAP estimator? What estimator minimizes the posterior expected loss?*