MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.s077 Spring 2018
Recitation 1 Friday 2/9

Topic 1. We first briefly reviewed the following useful properties, from probability class:

- Law of total probability: If the events $A_1, A_2, \ldots, A_n$ form a partition (of the sample space), that is, $\bigcup_{i=1}^n A_i = \Omega$, and, $A_i \cap A_j = \emptyset$, if, $i \neq j$; then,

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i).$$

  Here, a partition should be thought of as a collection, where, each outcome of the experiment belongs to a one and only one of them.

- Law of total expectation: Similar setup as above, if, X is a random variable, then,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i]\mathbb{P}(A_i).$$

  We then stated that, in case, there is another random variable, say, Y, taking values from a set, say, $\{1, 2, 3\}$, we could consider a partition:
  $A_i = \{Y = i\}$, for $i = 1, 2, 3$.
  This is also connected to the object, $\mathbb{E}[X \mid Y]$; which is a random variable, taking a value of $\mathbb{E}[X \mid Y = y]$; with probability $\mathbb{P}(Y = y)$. It is important to distinguish that, while, $\mathbb{E}[X \mid Y]$ is a random variable, $\mathbb{E}[X \mid Y]$ is a number.

Topic 2. We then moved on discussion about $\hat{P}$, and, $\hat{\mathbb{E}}$; where, the first is the empirical distribution, and, the second is the expectation, with respect to the empirical distribution.

- Let our setup be, $X_1, \ldots, X_n$ are i.i.d. random variables, with the observed values, $x_1, \ldots, x_n$, and if, $\hat{P}$ is their empirical distribution, where, X_i's take values from an alphabet, $\{a_1, \ldots, a_k\}$. Certain properties of $\hat{P}$ are as follows.

  - Let us begin by a simple warm-up. Let, $X_1, X_2, \ldots, X_n$ take values from the set, $\{0, 1\}$; and suppose that, we are to construct empirical distribution, corresponding to the observation, $(x_1, x_2, \dots, x_n)$. All possibilities are,

$$\left\{(0,1)\,, \left(\frac{1}{n}, \frac{n-1}{n}\right), \left(\frac{2}{n}, \frac{n-2}{n}\right), \dots, \left(\frac{n-1}{n}, \frac{1}{n}\right), (1,0)\right\}.$$

– Different observations can yield the same empirical distribution. For instance, following the setup as above, the observations, $(0, 1, 1, 0, 0, 1, 0)$, and, $(1, 1, 0, 0, 0, 0, 1)$ both have the empirical distribution, $(4/7, 3/7)$.

– If, $\hat{P} = (\hat{P}_1, \dots, \hat{P}_k)$, then, $n\hat{P} = (n\hat{P}_1, n\hat{P}_2, \dots, n\hat{P}_k)$ consists of k integers. Hence, the empirical distributions have a certain structure that, only some possible values are allowed. So, we cannot, for instance, have an empiric distribution with components, say, $(1/2, 1/\sqrt{10}, 1/2 - 1/\sqrt{10})$.

– We then stated that, $\hat{P}$ is a proxy for the actual, unknown distribution, from which we are obtaining i.i.d. samples. In fact, it converges to the actual distribution, in a certain sense (to be made precise in the problem set), as the number of samples goes to infinity.

– We also stated that, this distribution assigns weights to the set, $\{a_1, a_2, \dots, a_k\}$ (the technical word, is, $\hat{P}$ is supported on this set), that add up to one (this part was proven, by expressing the $\hat{P}_i$, as sum of indicators, and swapping a sum). Hence, it is a valid distribution, so, one can naturally consider $\hat{\mathbb{E}}$, that is, the expectation with respect to $\hat{P}$.

– **(This was briefly mentioned, for the sake of completeness, and is optional.)** Over the alphabet, $A = \{a_1, \dots, a_k\}$, let, $Q(A, n)$ be the total number of empirical distributions, corresponding to n observations. Very coarsely, if, $p \in Q(A, n)$, then, $p = (p_1, \dots, p_k)$, where, for each p_i, the numerator can take a value, from the set, $\{0, 1, \dots, n\}$, with denominator being n. Therefore, there is a total of, at most

$$|Q(A, n)| \le (n+1)^k,$$

empirical distributions. Hence, the total number of empirical distributions is polynomial in n. Of course, one can refine this analysis, and can come up with finer bounds, but this is not a very important detail.

• We next moved to discussion on $\hat{\mathbb{E}}$. As we have previously stated, since, $\hat{P}$ is a valid probability distribution, one can naturally ask the question: "What is the expected value, with respect to the empriic" We stated that We have noted in the lecture that, if, $X_1, \dots, X_n$ are i.i.d. random variables, with the observed values, $x_1, \ldots, x_n$, and if, $\hat{P}$ is their empirical distribution, we have,

$$\hat{\mathbb{E}}[X] = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

This is intuitively sound, and will be justified in the following lines. Let, X be a random variable, as above (taking values from an alphabet $\{a_1, a_2, \ldots, a_k\}$); and let, $X_1, \ldots, X_n$ be i.i.d. random variables; distributed according to X; and finally, $x_1, \ldots, x_n$ be its observed values. In the sequel, let, $I_{j,i}$ be a random variable, that takes a value 1, if, $X_j = a_i$; and 0, otherwise. Note that,

$$X_j = \sum_{i=1}^{k} I_{j,i} a_i.$$

$$\begin{aligned}
\hat{\mathbb{E}}[X] &= \sum_{i=1}^{k} \hat{P}_i a_i \\
&= \sum_{i=1}^{k} \sum_{j=1}^{n} \frac{1}{n} I_{j,i} a_i \\
&= \frac{1}{n} \sum_{j=1}^{n} \underbrace{\sum_{i=1}^{k} I_{j,i} a_i}_{=X_j} \\
&= \frac{1}{n} \sum_{j=1}^{n} X_j,
\end{aligned}$$

as desired.

Topic 3. Consider two random variables Y and Z, and a random variable X that is equal to Y with probability p and to Z with probability $1-p$.

1. Find the CDF of X in terms of the CDFs of Y and Z.
2. Assuming that both Y and Z are continuous, find the PDF of X in terms of the PDFs of Y and Z.
3. Find the mean of X in terms of the means of Y and Z
4. Find the variance of X in terms of the means and variances of Y and Z.

5. Suppose that Y and Z are both exponentially distributed, with parameters α and β. Formulate the optimization problem involved in the ML estimation of $\theta = (p, \alpha, \beta)$, given n i.i.d. samples, $x_1, \ldots, x_n$, drawn from the distribution of X.

6. For the same setting as in the previous part, but assuming that $p = 1/2$, follow the method of moments and write down a system of two equations whose solution will give an estimate of (α, β). Does this system have a unique solution?

Solution: Let, I be a random variable, which takes a value of 1; whenever $X = Y$ (hence, $\mathbb{P}(I = 1) = \mathbb{P}(X = Y) = p$); and a value 0, whenever, $X = Z$ (thus, $\mathbb{P}(I = 0) = \mathbb{P}(X = Y) = p$).

1. To compute $F_X(x) = \mathbb{P}(X \leq x)$; we condition on I; using law of total probability, as follows:

$$
\begin{aligned}
\mathbb{P}(X \leq x) &= \mathbb{P}(X \leq x | I = 1)\mathbb{P}(I = 1) + \mathbb{P}(X \leq x | I = 0)\mathbb{P}(I = 0) \\
&= \mathbb{P}(Y \leq x)p + \mathbb{P}(Z \leq x)(1 - p) \\
&= pF_Y(x) + (1 - p)F_Z(x).
\end{aligned}
$$

2. Since the PDF can be obtained, through differentiating the CDF, with respect to x, we have,

$$
f_X(x) = \frac{dF_X(x)}{dx} = p\frac{dF_Y(x)}{dx} + (1-p)\frac{dF_Z(x)}{dx} = pf_Y(x) + (1-p)f_Z(x).
$$

Here, there were a few questions about how a CDF looks like, those were clarified.

3. In this part, we use the law of total expectation.

$$
\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X | I = 1]\mathbb{P}(I = 1) + \mathbb{E}[X | I = 0]\mathbb{P}(I = 0) \\
&= \mathbb{E}[Y]p + \mathbb{E}[Z](1 - p) \\
&= p\mu_X + (1 - p)\mu_Y.
\end{aligned}
$$

4. For this part, we are asked to compute,

$$
\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.
$$

The second piece above, is already computed. Let us now, compute the first section.

$$\begin{aligned}\mathbb{E}[X^2] &= \mathbb{E}[X^2|I=1]\mathbb{P}(I=1) + \mathbb{E}[X^2|I=0]\mathbb{P}(I=0)\\ &= \mathbb{E}[Y^2]p + \mathbb{E}[Z^2](1-p)\\ &= p(\mu_Y^2 + \sigma_Y^2) + (1-p)(\mu_Z^2 + \sigma_Z^2),\end{aligned}$$

where, μ_y, μ_z are the means of Y, Z; and, σ_Y^2, σ_Z^2, are the variances of Y, Z. Summing up, the answer is,

$$p(\mu_Y^2 + \sigma_Y^2) + (1-p)(\mu_Z^2 + \sigma_Z^2) - (p\mu_Y + (1-p)\mu_Z)^2.$$

Above, we have noted the following. Often times, probability distributions are stated, with their mean, μ, and, the standard deviation, σ. Using this, you may need to compute the second moment as follows. First, let, T be a random variable, with the parameters above.

$$\mathrm{var}(T) = \sigma^2 = \mathbb{E}[T^2] - (\mathbb{E}[T])^2 = \mathbb{E}[T^2] - \mu^2 \implies \mathbb{E}[T^2] = \mu^2 + \sigma^2.$$

5. We first write, $f_Y(y) = \alpha e^{-\alpha y}$; and, $f_Z(z) = \beta e^{-\beta z}$. Now, using, $f_X = pf_Y + (1-p)f_Z$, we have,

$$f_X(x) = p\alpha e^{-\alpha x} + (1-p)\beta e^{-\beta x}.$$

Next, we write the likelihood function, for the observed sample, $(x_1, \ldots, x_n)$, with the help of independence:

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) = \prod_{i=1}^{n}(p\alpha e^{-\alpha x_i} + (1-p)\beta e^{-\beta x_i}).$$

Note that, even though we have not explicitly put the symbols, p, α, β on f, this likelihood depends on those parameters, as can be seen from the right-hand-side. Hence, the optimization procedure becomes,

$$\begin{aligned}(\hat{p}, \hat{\alpha}_{ML}, \hat{\beta}_{ML}) &= \text{argmax}_{p,\alpha,\beta} \prod_{i=1}^{n}(p\alpha e^{-\alpha x_i} + (1-p)\beta e^{-\beta x_i})\\ &= \text{argmax}_{p,\alpha,\beta} \sum_{i=1}^{n} \log(p\alpha e^{-\alpha x_i} + (1-p)\beta e^{-\beta x_i});\end{aligned}$$

where, the last equality holds, since, $\log(\cdot)$ is a monotonically increasing function. The last idea, is the fact that, optimizing an objective, which is expressed as a product, is often a hassle. However, taking log's convert the product into sum, where, one can simply take partial derivatives, with respect to p, α, β, set them equal to 0, and solve for the corresponding ML estimates.

6. Since we are to deal with two unknowns, namely, α, β, we need to have two equations.
Let us use, first and second empirical moments. First, recall that, if $X_1, \dots, X_n$ are i.i.d., with the observed values, $x_1, \dots, x_n$, then,

$$\hat{m}_1 = \hat{\mathbb{E}}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

where, $\hat{\mathbb{E}}[\cdot]$ denotes the expectation, with respect to the empirical mean. Now, we have the first equation:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \frac{p}{\alpha} + \frac{1-p}{\beta} = \frac{1}{2\alpha} + \frac{1}{2\beta}.$$

We can similarly, cast a same system for the second moment. Recall that, for an exponential distribution Z, with parameter λ, $\mathbb{E}[Z^2] = \frac{2}{\lambda^2}$. With this, the equation for the second moment estimate, becomes,

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{2\alpha^2} + \frac{1}{2\beta^2}.$$

This is a system with two equations, involving two unknowns (namely, α, β), and can be solved. For the sake of completeness, a sketch is below. Let, $S_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$, and, S_2 similarly denote, $S_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$. We have,

$$2S_1 = \frac{1}{\alpha} + \frac{1}{\beta} \implies 4S_1^2 = \underbrace{\frac{1}{\alpha^2} + \frac{1}{\beta^2}}_{=2S_2} + \frac{2}{\alpha\beta} \implies \frac{1}{\alpha\beta} = 2S_1^2 - S_2.$$

Letting, $1/\beta = \alpha(2S_1^2 - S_2)$, we arrive at,

$$2S_1 = \frac{1}{\alpha} + \alpha(2S_1^2 - S_2),$$

which is a quadratic in α.
Next, note that, since the expressions above are symmetric in α, β, we actually get two solutions, (α, β), and, (β, α). This is a consequence of the fact that, $p = 1/2$. However, the solution here is unique, up to permutations (which can be seen that, quadratic has two possibilities, and whichever possibility is selected, the remaining is the value of β).