MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.s077 Spring 2018

Partial Lecture 2 Notes and **Problem Set 1** due Thursday 2/15, in class

Readings: Slides from Lectures 2. We do not provide detailed notes for Lecture 2. However, the problem statements below do sketch many of the concepts in Lecture 2 in a self-contained manner, and can be viewed as a lecture-note/problem-set combination.

Problem 1. The relation between true and empirical distributions.

This is a theoretical problem, which is actually easy once you refresh your understanding of the weak law of large numbers.

Let $X, X_1, \ldots, X_n$ be i.i.d. random variables. We use the symbol $\mathbb{P}$ to refer to probabilities that conform to the distributions of these random variables.

We observe realized values $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$, and form the **empirical distribution** $\widehat{\mathbb{P}}$. In particular, for any set A

$$\widehat{\mathbb{P}}(A) = \frac{\text{number of samples that belong to } A}{n}.$$

Another way of thinking about the empirical distribution is that it is a discrete distribution that when the x_i are all distinct, assigns equal probability $(1/n)$ to each observed value x_i.

Note that $\widehat{\mathbb{P}}$ can be viewed in two ways:

1. It is a legitimate probability distribution (nonnegative weights that add to 1).
2. It is a random object, because it is determined by the random realizations of $X_1, \ldots, X_n$.

We will be using the notation $\widehat{\mathbb{E}}[\cdot]$ to denote expectations calculated with respect to the empirical distribution $\widehat{\mathbb{P}}$.

A wide variety of statistical methods are based on the assumption that $\widehat{\mathbb{P}}_n$ is close to $\mathbb{P}$. This is not guaranteed with certainty: it is possible that the realized values of the X_i are all outliers, not representative of the distribution $\mathbb{P}$. On the other hand, there is high probability that $\widehat{\mathbb{P}}$ will indeed be close to $\mathbb{P}$. The purpose of this exercise is to provide some evidence in this direction.

Suppose that X is a discrete random variable, taking values in a finite set $\{a_1, \ldots, a_k\}$, and let $p_i = \mathbb{P}(X = a_i)$. The distribution $\mathbb{P}$ is completely determined by the vector $p = (p_1, \ldots, p_k)$. Similarly, the empirical distribution $\widehat{\mathbb{P}}$ is completely determined by $(\widehat{P}_1, \ldots, \widehat{P}_k)$, where

$$\widehat{P}_i = \frac{\text{number of } j \text{ for which } X_j = a_i}{n}, \qquad i = 1, \ldots, k,$$

is the **empirical frequency** of outcome a_i. Note that each $\widehat{P}_i$ is a random variable. In particular $n\widehat{P}_i$ is binomial with parameters n and p_i. Furthermore, the weak law of large numbers tells us that $\widehat{P}_i$ converges to p_i , in the following sense ("in probability"). Fix any $\epsilon > 0$. Then,

$$\lim_{n\to\infty} \mathbb{P}\Big(|\widehat{P}_i - p_i| \leq \epsilon\Big) = 1, \qquad \text{for all } i.$$

At this point you should review the derivation of the above statement, using the fact that the variance of $\widehat{P}_i$ goes to zero and the Chebyshev inequality.

(a) We now argue that the **vector** $\widehat{P}$ (that is all entries $\widehat{P}_i$, simultaneously) converges to the vector p.

Fix again some $\epsilon > 0$. Show that

$$\lim_{n\to\infty} \mathbb{P}\Big(|\widehat{P}_i - p_i| \leq \epsilon, \text{ for all } i\Big) = 1.$$

Hint: Use the "union bound."

(b) Suppose now that X is a general random variable. Let $\mathbb{F}(\cdot)$ be the CDF of X, and let $\widehat{F}$ be the empirical CDF:

$$\widehat{F}(x) = \frac{\text{number of } j \text{ for which } X_j \leq x}{n}.$$

Fix some $\epsilon > 0$. Show that, for any given $x \in \mathbb{R}$,

$$\mathbb{P}\Big(|\widehat{F}(x) - \mathbb{F}(x)| > \epsilon\Big)$$

converges to zero as $n \to \infty$.

Deep dive: More information on the relation between $\mathbb{P}$ and $\widehat{\mathbb{P}}$. Part (b) of the preceding problem establishes, that for any fixed x, $\widehat{F}(x)$ converges to $\mathbb{F}(x)$, as n increases. In principle, this convergence could be slower and slower, as we consider more extreme values of x. However, this won't happen: convergence takes place simultaneously, for all x; the mathematical term is "uniform convergence." The mathematical statement, known as the *Glivenko-Cantelli Theorem* asserts that

$$\mathbb{P}\Big(\max_x |\widehat{F}(x) - \mathbb{F}(x)| > \epsilon\Big)$$

converges to zero as $n \to \infty$. That is, when n is large, and with high probability, $\widehat{F}$ is everywhere close to $\mathbb{F}(x)$. In fact, the same is true under a stronger notion of convergence ("almost sure convergence" or "convergence with probability 1," in case you are familiar with this notion). We will not touch the fairly advanced proof of the Glivenko-Cantelli theorem, but it is a most useful fact to know.

The plugin method for estimation. The convergence of the empirical distribution to the true distribution provides the conceptual foundation behind **plugin methods**. We are interested in some quantity α that describes a property of a distribution $\mathbb{P}$, according to some formula; abstractly speaking, $\alpha = h(\mathbb{P})$, for a function h that takes as input a probability distribution and produces a real number. Some examples: $h(\mathbb{P}) = \mathbb{P}(X \geq 0)$ or $h(\mathbb{P}) = \mathbb{E}[X^4]$. In mathematical parlance, h or α is called a **functional** of the distribution $\mathbb{P}$.

Suppose now that $\mathbb{P}$ is available, but we have access to samples drawn independently according to $\mathbb{P}$. On the basis of these samples we can construct the empirical distribution $\widehat{\mathbb{P}}$, and estimate α by letting $\hat{\alpha} = h(\widehat{\mathbb{P}})$.

Problem 2. The purpose of this problem is to illustrate various estimation methods, for a very simple distribution.

Let $X, X_1, \ldots, X_n$ be i.i.d. nonnegative random variables, drawn from an exponential distribution with parameter $\theta > 0$, with corresponding PDF given by

$$f^{\theta}_{X_i}(x) = \theta e^{-\theta x}, \qquad \forall\, x \geq 0.$$

Recall that $\mathbb{E}[X_i] = 1/\theta$ and $\mathrm{var}(X_i) = 1/\theta^2$. We wish to estimate θ on the basis of observed values $x_1, \ldots, x_n$ of $X_1, \ldots, X_n$.

(a) Write down the form of the **likelihood function**

$$f^{\theta}_{X_1,\ldots,X_n}(x_1, \ldots, x_n),$$

take its logarithm, and then set its derivative to zero to maximize it. Show that the Maximum Likelihood (ML) estimate (i.e., the estimate that maximizes the above likelihood function over θ) is of the form

$$\hat{\theta}_{\mathrm{ML}} = \frac{n}{x_1 + \cdots + x_n}.$$

The method of moments. This method first estimates some of the moments of a distribution using the plugin methodology, and uses these estimates to solve for the unknown parameters of interest. As a concrete instance, the method of moments based on the kth moment finds an estimate $\hat{\theta}$ by finding a solution (for θ) of the equation $\widehat{\mathbb{E}}[X^k] = \mathbb{E}^{\theta}[X^k]$. Here, we use the superscript θ to make it explicit that the distribution, and therefore expectations as well, are affected by θ.

(b) Use the method of moments with $k = 1$, and then with $k = 2$, to derive two different estimates of θ, to be denoted by $\hat{\theta}_1$ and $\hat{\theta}_2$, respectively.

(c) Let a_θ be the median of the distribution of X; that is,

$$\mathbb{P}(X \leq a_\theta) = \frac{1}{2}.$$

Find a closed-form expression for a_θ. Let $\hat{a}$ be the median of the empirical distribution. (Assume that n is odd, so that the median is uniquely defined; in particular, once we sort the data x_i, then $\hat{a}$ is the "middle" value.)

Find a formula (in terms of $\hat{a}$) for the **feature-matching** estimate $\hat{\theta}_m$ that uses the median as a feature. That is, we let $\hat{\theta}_m$ be the solution, for θ, of the equation $a_\theta = \hat{a}$.

(d) You are given a data set with $n = 50$ samples, in a data file (data_2d.csv) provided on the computation portal. Your task is to calculate the numerical values of the four estimates introduced in parts (a)-(c) of this problem. You should submit the .ipynb file with the code and numerical estimates along with your solution write-up. We provide a skeleton .ipynb file on the portal. You can access it as follows:

- Log on to the portal (`http://mit-6s077.mit.edu`)
- After you have successfully logged in, in the same browser window, copy and paste the following assignment url: (`http://mit-6s077.mit.edu/hub/user-redirect/git-sync?repo=https://github.com/jehangiramjad/mit-6s077&subPath=Pset1/Problem_2d.ipynb)`)

Problem 3. The purpose of this problem is twofold:

(i) To explore the application of the methods we have introduce to the problem of learning the parameters of a mixture of distributions;

(ii) to work on the Naples crab data set introduced in lecture. You are given the dataset in a data file (crab_data.csv) provided on the computation portal. Your task is to use the dataset to produce the estimates and plots asked in parts (b), (e) and (h) of this problem. You should submit the .ipynb file with the code and numerical estimates along with your solution write-up. We provide a skeleton .ipynb file on the portal. You can access it as follows:

- Log on to the portal (http://mit-6s077.mit.edu)
- After you have successfully logged in, in the same browser window, copy and paste the following assignment url: (http://mit-6s077.mit.edu/hub/user-redirect/git-sync?repo=https://github.com/jehangiramjad/mit-6s077&subPath=Pset1/Problem_3.ipynb)

We start by modeling the data set as if each data point x_i is the realized value of a normal random variable with unknown mean μ and variance v. The likelihood function is of the form

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi v}} \exp\Big\{ -(x_i-\mu)^2/2v \Big\}.$$

(a) By taking logarithms and then setting the derivative to zero, verify that the ML estimate is of the form

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \hat{v} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2.$$

(b) For the crab data set, calculate the numerical value of the ML estimate of the 2-dimensional parameter vector $\theta = (\mu, v)$. Plot the estimated PDF (a normal PDF whose parameters are set equal to the ML estimates) and compare visually with the histogram of the original data.
You should do this using the computation platform. Access instructions to the data and the skeleton .ipynb file are provided above.

Quantifying the goodness of fit between an estimated distribution and the data. We use the data to come up with an estimate $\hat{\theta}$ of the parameter vector. This estimate results in an estimated distribution; let $\mathbb{F}^{\hat{\theta}}$ be the CDF of the estimated distribution. If estimation is "successful," then $\mathbb{F}^{\hat{\theta}}$ should be fairly "close" to the empirical CDF $\widehat{F}$. Thus, a reasonable measure of the *goodness of fit* is

$$\delta(\hat{\theta}) = \max_x |F^{\hat{\theta}}(x) - \hat{F}(x)|.$$

For the Naples crab example, and with the above considered estimation procedure (a single normal), it turns out that $\delta(\hat{\theta}) = 0.06029$.

Mixture models. Visual inspection of the histogram in part (b) of this problem suggests that the underlying model is more complex. Perhaps the crab population consists of two different species, with respective proportions λ and $1-\lambda$, each species being described by a different normal distribution, with its own mean and variance, μ_i and v_i, $i = 1, 2$. If so, we have a 5-dimensional parameter vector θ to be estimated, namely, $\theta = (\lambda, \mu_1, v_1, \mu_2, v_2)$.

(c) Write down a formula for the density $f^{\theta}_{X_i}(x)$ of X_i, under a given parameter vector θ.

Maximum likelihood estimation of mixture models. The likelihood function is of the form

$$L^{\theta}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f^{\theta}_{X_i}(x_i).$$

Because of the form of the answer to part (c) above, this is a rather complicated nonlinear, nonquadratic, nonconvex function of θ, and its maximum (over θ) cannot be found in closed form. Instead one must resort to numerical methods. Such numerical methods tend to be of two types:

(i) General purpose hill climbing: starting with some initial θ, you consider the gradient of L^{θ} with respect to θ (and perhaps its second derivatives as well) and move to a nearby θ at which the likelihood is higher, until a local maximum is reached.

(ii) Special purpose methods, with the so-called "EM algorithm" being the most popular one. Such algorithms are beyond the scope of this course. Suffice to say that one can rely on black-box implementations available in many software libraries. For example, the Gaussian Mixture Library in Scikit-Learn (`http://scikit-learn.org/stable/modules/mixture.html`), can be used to estimate the Gaussian Mixture using the EM algorithm.

The above mentioned numerical methods unfortunately do not have any strong guarantees. The likelihood function may have multiple local maxima; if the algorithm is intialized in the vicinity of a local maximum, it may get "stuck" there and never identify the true maximum.

(d) Consider an estimate of θ in which λ is set to 1/2, μ_1 is set to x_1 (the first data point), and v_1 is set to a small positive number. (The other parameters, μ_2 and v_2 are set to some arbitrary, fixed values. What happens to the likelihood function as v_1 goes to zero? What is the maximum value that L^{θ} will reach, when we optimize over θ?

Maximum likelihood for learning mixture models needs care. The answer to part (d) indicates that if the variance estimates are allowed to be arbitrarily small, the maximum likelihood method will produce unsound estimates. In practice, this can be taken care of by constraining v_i to be larger than a positive constant. This tends to eliminate spurious, undesirable solutions.

For the Naples crab example, after taking care to eliminate such unsound solutions, we obtain the maximum likelihood estimates using the EM-algorithm provided by Scikit-Learn (referenced above):

$$\hat{\lambda} = 0.648536, \qquad \hat{\mu}_1 = 0.656405, \qquad \hat{v}_1 = 0.000156,$$

$$\hat{\mu}_2 = 0.628942, \qquad \hat{v}_2 = 0.000256.$$

The corresponding value of the goodness of fit measure $\delta(\hat{\theta})$ turns out to be equal to 0.01785.

(e) Plot the estimated PDF, for the mixture model whose parameters are set equal to the above ML estimates, and compare visually with the histogram of the original data.
You should do this using the computation platform. Access instructions to the data and the skeleton .ipynb file are provided above.

It should not be a surprise that by using a mixture model, with more parameters, we can fit the data better, as evidenced both visually and from the goodness of fit measure. A mixture of three (or four, etc.) normals would do even better. Where do we stop? At this point, other than some visual evidence we do not really have a methodology for declaring with some confidence that there is only one (versus two, or three) species. We are dealing here with a *hypothesis testing* problem. Such problems will be studied in more detail later in the course.

Using the method of moments. An alternative approach to estimating the 5-dimensional parameter vector is provided by the method of moments. Let $m_k(\theta)$ be the kth moment, $\mathbb{E}[X^k]$, under a particular parameter vector θ. Let $\hat{m}_k$ be the kth moment of the empirical distribution of X. In the current context, the method of moments estimates the parameter vector by solving the following system of equations (5 equations in 5 unknowns)

$$m_k(\theta) = \hat{m}_k, \qquad k = 1, \ldots, 5.$$

This is a pretty complicated system of equations, which can only be solved numerically. In the next two parts, we go over what it takes to set up this system of equations.

(f) Use the data to calculate the numerical values of $\hat{m}_k$, $k = 1, \ldots, 5$.
You should do this using the computation platform. Access instructions to the data and the skeleton .ipynb file are provided above.

(g) For a standard normal random variable Z, the symmetry of the distribution yields $\mathbb{E}[Z] = \mathbb{E}[Z^3] = \mathbb{E}[Z^5] = 0$. Furthermore, $\mathbb{E}[Z^2] = 1$. Finally, $\mathbb{E}[Z^4] = 3$, as can be found through integration by parts. If Y is a nornal random variable with mean μ and variance v, express $\mathbb{E}[Y^2]$ and $\mathbb{E}[Y^3]$ as a function of μ and v. We spare you the calculation of the other moments, which happen to be

$$\mathbb{E}[Y^4] = \mu^4 + 6\mu^2 v + 3v^2, \qquad \mathbb{E}[Y^5] = \mu^5 + 10\mu^3 v + 15\mu v^2.$$

Given the above facts, use the total expectation theorem to express the fifth moment of X, when X is described by our mixture model, as a function of $\lambda, \mu_1, v_1, \mu_2, v_2$.

Armed with the numerical values of the $\hat{m}_k$ and the closed form expressions for the $m_k(\theta)$, the system of equations can be written down. We run it through a numerical solver and obtain

$$\hat{\lambda} = 0.351, \qquad \hat{\mu}_1 = 0.629, \qquad \hat{v}_1 = 0.000256 \qquad ,$$

$$\hat{\mu}_2 = 0.656, \qquad \hat{v}_2 = 0.000156.$$

The corresponding value of the goodness of fit measure $\delta(\hat{\theta})$ turns out to be equal to 0.01789. Notice that the estimates are pretty much the same as those provided by the EM algorithm discussed earlier–this provides a good validation for the fact that when feasible these approaches are both valid.

(h) (Open-ended) After going through all of the above calculations, suppose that we have been convinced that X is indeed described by a mixture of two (as opposed to one) normals. Does this mean that we are faced with two different species? Are there alternative possible explanations? (Be imaginative and creative.)