EE2211 : Introduction to Machine Learning (Fall 2020)
A Crash Course on Probability and Maximum Likelihood Estimation

Vincent Y. F. Tan

In this brief document, we will summarize some key concepts from probability and provide some examples of maximum likelihood estimation. The topics of probability, statistics, and estimation theory are so rich and vast that a few pages will not do justice to them. Nevertheless, we will try.

A word about notation. Random variables and the values that they take on will be denoted in upper (e.g., X) and lower case (e.g., x, x_i, x) respectively. Sets (in which the random variables assume their values) will be denoted by calligraphic font (e.g., $\mathcal{X}$). The probability, expectation and variance operators are respectively denoted by $\Pr(\,\cdot\,)$, $\mathbb{E}[\,\cdot\,]$ and $\mathrm{Var}(\,\cdot\,)$.

1 Basic Probability Theory

Because all of modern statistical machine learning deals with uncertainty, it seems appropriate to start of by reminding ourselves of the definitions of events, probabilities, joint probabilities and conditional probabilities. For more details, the reader is encouraged to consult any standard probablity textbook such as those by Bertsekas and Tsitsiklis [BT02] or Ross [Ros12].

Let us motivate probability and statistical (Bayesian) inference by considering an example from Bishop's machine learning book [Bis08].

Example 1. *There is a red box and a blue box. In the red box there are a total of 8 fruits, 2 apples and 6 oranges. In the blue box, there are a total of 4 fruits, 3 apples and 1 orange. The probability of selecting the red box is* $\Pr(B = \mathsf{r}) = 2/5$ *and probability of selecting the blue box is* $\Pr(B = \mathsf{b}) = 3/5$*. Having selected a box, selecting any item within the box is equally likely. Some of the questions we would like to ask include:*

- *What's the probability that we select an orange?*
- *Given that we've selected an orange, what's the probability that we chose it from the blue box?*

1.1 Joint and Conditional Probabilities for Discrete Random Variables

Now, let us consider a more general example involving two random variables X and Y. Suppose, as in the example above, the random variables are only permitted to take on *finitely many* values. So X can only take on values in the finite set $\mathcal{X} = \{x_1, \ldots, x_M\}$ and Y takes on values in $\mathcal{Y} = \{y_1, \ldots, y_L\}$. Such random variables are known as *discrete* random variables (in which no mathematical peculiarities arise). It should be clear that

$$\sum_{i=1}^{M} \Pr(X = x_i) = 1. \tag{1}$$

Consider n trials (of sampling X and Y) and let the number of trials for which $X = x_i$ and $Y = y_j$ be n_{ij}. Then, if n is large, we can assume that

$$\Pr(X = x_i, Y = y_j) = \frac{n_{ij}}{n}. \tag{2}$$

The argument of $\Pr(\cdot)$ is known as an *event*. Roughly speaking, an event is a set in which a random variable (or multiple random variables) assume some value(s) in some set, e.g., $\{X = x_i, Y = y_j\}$. What's the probability that $X = x_i$? Well, we simply sum up those n_{ij}'s for which the first index equals to i. In other words,

$$\Pr(X = x_i) = \sum_{j=1}^{L} \frac{n_{ij}}{n} =: \frac{c_i}{n}, \tag{3}$$

where $c_i := \sum_{j=1}^{L} n_{ij}$. Expressed slightly differently, we have

$$\Pr(X = x_i) = \sum_{j=1}^{L} \Pr(X = x_i, Y = y_j). \tag{4}$$

This is the important *sum rule* in probability and will be used extensively in statistical machine learning so the reader is urged to internalize this. Similarly, the marginal probability that $Y = y_j$ is

$$\Pr(Y = y_j) = \sum_{i=1}^{M} \Pr(X = x_i, Y = y_j). \tag{5}$$

Now, we introduce the important notion of conditional probabilities. Given that $X = x_i$, what is the probability that $Y = y_j$? Clearly,

$$\Pr(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}. \tag{6}$$

But note also that

$$\Pr(X = x_i, Y = y_j) = \frac{n_{ij}}{n} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{n} = \Pr(Y = y_j \mid X = x_i) \Pr(X = x_i). \tag{7}$$

We have derived the important *product-rule* in probability. The (two basic) rules of probability are summarized as follows:

$$\Pr(X = x) = \sum_{y \in \mathcal{Y}} \Pr(X = x, Y = y), \tag{8}$$

$$\Pr(X = x, Y = y) = \Pr(Y = y \mid X = x) \Pr(X = x). \tag{9}$$

A note about notation. We will usually write $p_X(x) := \Pr(X = x)$ or simply denote this function as $p(x)$ when the random variable is clear from the context. The function $p(x)$ is known as the *probability mass function* or pmf and it satisfies

$$p(x) \geq 0 \quad \forall\, x \in \mathcal{X} \qquad \text{and} \qquad \sum_{x \in \mathcal{X}} p(x) = 1. \tag{10}$$

You will also often hear the colloquial term "distribution". This is often used synonymously with "pmf". Similarly, the joint pmf of random variables X and Y is denoted as $p(x, y) := \Pr(X = x, Y = y)$. Finally, the conditional will be denoted as $p(y \mid x) := \Pr(Y = y \mid X = x)$.

By combining the sum rule in (8) and the product rule in (9), we can derive Bayes' rule:

$$p(y \mid x) = \frac{p(x, y)}{p(x)} = \frac{p(x \mid y)p(y)}{p(x)} = \frac{p(x \mid y)p(y)}{\sum_{y' \in \mathcal{Y}} p(x \mid y')p(y')}. \tag{11}$$

This is a central relationship in pattern recognition, machine learning and statistical physics. Note that we have "inverted the causal relationship" between X and Y. On the left, Y "depends on" X while on the right, we have expressed the same causality relationship in terms of the causal dependence of X on Y.

Bayes' theorem can be written alternatively as follows:

$$p(x \mid y) \propto p(y \mid x)p(x), \tag{12}$$

where $\propto$ denotes equality up to a constant (not depending on x). If x designates an unknown variable, something we would like to infer and $p(x)$ is its *prior probability*, then $p(x \mid y)$ denotes the *posterior probability*, the belief we have about x *after*[1] we know that $Y = y$. In the parlance of statistical inference, Bayes' rule in (11) can be written as

$$\text{posterior } = \frac{\text{likelihood } \times \text{prior}}{\text{model evidence}} \propto \text{likelihood } \times \text{prior } . \tag{13}$$

We now use this relation to solve:

Exercise 1. *Let F be the random variable denoting the fruit chosen. Using the sum, product and Bayes' rules, verify that the answers to the questions in Example 1 are* $\Pr(F = \mathsf{o}) = 9/20$ *and* $\Pr(B = \mathsf{b} \mid F = \mathsf{o}) = 1/3$ *respectively.*

Note the following: Prior to having any additional information about what fruit we chose, the *prior probability* of choosing from a blue box is $\Pr(B = \mathsf{b}) = 3/5$. However, if we know the identity of the fruit we chose, say orange, then the *posterior probability* of choosing from the blue box is $\Pr(B = \mathsf{b} \mid F = \mathsf{o}) = 1/3$. This is the simplest non-trivial example of statistical inference. Intuitively, this is true because the blue box contains far fewer oranges so knowing that we chose an orange biases our *belief* about the box we chose from.

1.2 Continuous Random Variables

We now segue into the land of continuous random variables. To introduce continuous random variables formally, we would require too much mathematical machinery that goes beyond the scope of the class; see [Ros12] for a gentle introduction. However, very roughly speaking, X is said to be a *continuous random variable* if its *cumulative distribution function* (cdf) $x \in \mathbb{R} \mapsto F_X(x) = \Pr(X \leq x)$ is differentiable on $\mathbb{R}$.[2] Its derivative is then called the *probability density function* (pdf) of X and written as

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x} F_X(x), \quad x \in \mathbb{R}. \tag{14}$$

Any pdf has the following properties (why?):

$$f_X(x) \geq 0 \quad \forall\, x \in \mathbb{R} \qquad \text{and} \qquad \int_{\mathbb{R}} f_X(x)\,\mathrm{d}x = 1. \tag{15}$$

Now, it is easy to see (cf. the second fundamental theorem of calculus) that for any $-\infty < a \leq b < \infty$,

$$\Pr(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,\mathrm{d}x. \tag{16}$$

We can define the *joint density* (or *joint probability density function*) of a pair of continuous random variables as

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x\, \partial y} F_{X,Y}(x, y), \quad (x, y) \in \mathbb{R}^2. \tag{17}$$

Conditional probability and independence can be defined analogously to the discrete case. Clearly,

$$\Pr(a < X \leq b, c < Y \leq d) = \int_c^d \int_a^b f_{X,Y}(x, y)\,\mathrm{d}x\,\mathrm{d}y. \tag{18}$$

[1] The word "posterior" is derived from the word "post", which means "after" (e.g., post-midnight means after midnight).

[2] Roughly speaking, a function $g : I \to \mathbb{R}$ defined on an interval $I \subset \mathbb{R}$ is said to be *differentiable at $x \in I$* if the limit $\lim_{\epsilon \to 0}(g(x + \epsilon) - g(x))/\epsilon$ exists. In addition, $g : I \to \mathbb{R}$ is said to be *differentiable on I* if it is differentiable at each $x \in I$.

1.3 Independence

What does it mean for two random variables X and Y to be independent? Roughly speaking, knowledge of one does not influence our knowledge of the other. More precisely, this can be expressed in a variety of ways. Two random variables X and Y are *independent* if their joint distribution $p_{X,Y}(x,y) := \Pr(X = x, Y = y)$ factorizes, i.e.,

$$p_{X,Y}(x,y) = p_X(x)p_Y(y), \tag{19}$$

for every $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Note from (19) that if X and Y are independent if

$$p_{X|Y}(x \mid y) = p_X(x) \tag{20}$$

for every $x \in \mathcal{X}$ and *every* $y \in \mathcal{Y}$ *such that* $p_Y(y) > 0$ (what if $p_Y(y) = 0$?). Intuitively, (20) means that knowledge that $Y = y$ tells you no additional information about X.

For example, I toss a coin $2n$ times and all tosses are mutually independent. Let X be the number of heads that I observe in the first n coin tosses and let Y be the number of heads that I observe in the second n coin tosses. Since X and Y are the result of independent coin tosses, the two random variables X and Y are independent. Consider another example. Let X and Y denote respectively, the presence (or absence) of rain in Jurong and in Clementi respectively. These random variables are clearly not independent because if I know that there is rain in Jurong, due to the proximity of the two locations, it is also likely that there is rain in Clementi. So knowledge of X *does* provide some information about Y.

1.4 Expectation and Variance

The *expectation* of a discrete random variable X with pmf p_X is defined to be

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x p_X(x). \tag{21}$$

If X is a continuous random variable with pdf f_X, we have

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x. \tag{22}$$

Note that the expectation is a statistical summary of the *distribution* of X, rather than depending on the realized value of X.

If g is a function from the domain of X to $\mathbb{R}$, we can obtain the expectation of $Y = g(X)$ in the same way. It can be shown that

$$\mathbb{E}[Y] = \mathbb{E}[g(X)] = \int_{\mathbb{R}} y f_Y(y)\,\mathrm{d}y = \int_{\mathbb{R}} g(x) f_X(x)\,\mathrm{d}x. \tag{23}$$

In particular if $g(X) = aX + b$ (for constants $a, b \in \mathbb{R}$), then $\mathbb{E}[g(X)] = a\mathbb{E}[X] + b = g(\mathbb{E}[X])$. This fact is known as the *linearity of expectation*.

The *variance* of X is the expectation of $g(X) = (X - \mathbb{E}[X])^2$, a particular function of X. Thus,

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 p_X(x) \qquad \text{(discrete rvs)}, \tag{24}$$

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \int_{\mathbb{R}} (x - \mathbb{E}[X])^2 f_X(x)\,\mathrm{d}x \qquad \text{(continuous rvs)}. \tag{25}$$

Exercise 2. *Check from the above definition that the variance can also be expressed as*

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{26}$$

2 Maximum Likelihood Estimation

In machine learning, we almost always do not have access to the underlying distributions that the data are generated from. Rather, we have access to *sample* data and we would like to use it to estimate some parameters. For example, in class, you saw that if $\mathcal{S} = \{X_1, \ldots, X_n\}$ are independent samples from the univariate Gaussian distribution

$$f_X(x; \mu, \sigma^2) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}(x-\mu)^2 \Big) \quad x \in \mathbb{R}, \tag{27}$$

then we can use the samples in $\mathcal{S}$ to estimate the parameter vector $\theta = (\mu, \sigma^2)$. Note that we use the notation $f_X(x; \mu, \sigma^2)$ or $f_X(x; \theta)$ to emphasize that the density is *parametrized* by the parameters (μ, σ^2) or θ.[3] In the following, I would like to expand on this point by providing you with a couple more examples.

Example 2 (Bernoulli Distribution)**.** *Let us say that the samples in $\mathcal{S} = \{X_1, \ldots, X_n\}$ are generated independently from the Bernoulli (coin toss) distribution*

$$p_X(x; \theta) = \begin{cases} 1-\theta & x = 0 \\ \theta & x = 1 \end{cases}. \tag{28}$$

It would be convenient to write this as

$$p_X(x; \theta) = (1-\theta)^{1-x}\theta^x, \quad x \in \{0, 1\}. \tag{29}$$

Check that this is true. The mean of the distribution $\mathbb{E}[X]$ is clearly $\theta \in (0, 1)$ (check). How would we estimate θ from samples? Consider,

$$\hat{\theta}_{\text{ML}} = \arg\max_{\theta \in (0,1)} \quad \prod_{i=1}^{n} p_X(X_i; \theta) \tag{30}$$

$$= \arg\max_{\theta \in (0,1)} \quad \sum_{i=1}^{n} \log p_X(X_i; \theta) \tag{31}$$

$$= \arg\max_{\theta \in (0,1)} \quad \sum_{i=1}^{n} [(1 - X_i)\log(1-\theta) + X_i \log\theta] \tag{32}$$

$$= \arg\max_{\theta \in (0,1)} \quad (n - N_1)\log(1-\theta) + N_1 \log\theta \tag{33}$$

where we have used $N_1 := \sum_{i=1}^{n} X_i$ to denote the total number of ones in $\mathcal{S}$. Since the objective function is strictly concave, differentiating and setting to zero yields the (unique) maximum which is

$$-\frac{n - N_1}{1 - \hat{\theta}_{\text{ML}}} + \frac{N_1}{\hat{\theta}_{\text{ML}}} = 0 \qquad \Longrightarrow \qquad \hat{\theta}_{\text{ML}} = \frac{N_1}{n}. \tag{34}$$

So the mean θ is estimated by using the empirical mean N_1/n, which agrees with common sense!

Example 3 (Exponential Distribution)**.** *Now we consider an example involving continuous random variables. Let us say that the samples in $\mathcal{S} = \{X_1, \ldots, X_n\}$ are generated independently from the exponential distribution (this distribution models waiting times for buses)*

$$f_X(x; \theta) = \begin{cases} \theta \exp(-\theta x) & x \geq 0 \\ 0 & x < 0 \end{cases}. \tag{35}$$

[3] Contrast this to the notation $f_{X|Y}(x \mid y)$ in which Y is *random* (and takes on the value y). When we use the semicolon in $f_X(x; \theta)$, the parameter θ is not random; rather it is *deterministic* but *unknown*.

Here, $\theta > 0$ is the unknown (rate) parameter. In fact, $1/\theta = \mathbb{E}[X]$ is the mean of the exponential distribution. Check this by integration by parts. How would we estimate θ from samples? Consider,

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta>0} \quad \prod_{i=1}^{n} f_X(X_i;\theta) \tag{36}$$

$$= \arg\max_{\theta>0} \quad \sum_{i=1}^{n} \log f_X(X_i;\theta) \tag{37}$$

$$= \arg\max_{\theta>0} \quad \sum_{i=1}^{n} (\log\theta - \theta X_i) \tag{38}$$

$$= \arg\max_{\theta>0} \quad n\log\theta - \theta\sum_{i=1}^{n} X_i. \tag{39}$$

Since the objective function is strictly concave, differentiating and setting to zero yields the (unique) maximum which is

$$\frac{n}{\hat{\theta}_{\mathrm{ML}}} = \sum_{i=1}^{n} X_i \qquad \Longrightarrow \qquad \hat{\theta}_{\mathrm{ML}} = \frac{n}{\sum_{i=1}^{n} X_i} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^{-1}. \tag{40}$$

Notice that the result makes sense because $\frac{1}{n}\sum_{i=1}^{n} X_i$ is the empirical mean. Indeed, since θ is the reciprocal of the mean $\mathbb{E}[X]$, it seems plausible that $n/(\sum_{i=1}^{n} X_i)$ is a "good" estimate of θ. Why would it not be "good"? Hint: What happens to $\mathbb{E}[\hat{\theta}_{\mathrm{ML}}]$ when n is small?

Exercise 3. *Assume that the samples in $\mathcal{S} = \{X_1, \ldots, X_n\}$ are sampled independently from the uniform distribution*

$$f_X(x;\theta) = \begin{cases} 1/\theta & 0 \le x \le \theta \\ 0 & else \end{cases}. \tag{41}$$

Find the maximum likelihood estimate of θ. Discuss the (peculiar) properties of the estimate.

References

[Bis08] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2008.

[BT02] D. P. Bertsekas and J. N. Tsitsiklis. *Introduction to Probability.* Athena Scientific, 1st edition, 2002.

[Ros12] S. Ross. *A First Course in Probability.* Pearson, 9th edition, 2012.