CIS3990-002: Mathematics of Machine Learning Fall 2023

Lecture: Probability

Date: September 6th, 2023 *Author: Eric Wong*

1 Probability Basics

- Probability space $(\Omega, \mathcal{A}, P)$ is a real-world process with random outcomes (i.e. an experiment). Ex. Flip two coins and see how many heads show up.
- Sample space Ω: set of all possible outcomes of the experiment. Ex. {hh, tt, ht, th}
- Event space $\mathcal{A}$: the space of potential results (events). Ex. the power set of of Ω.
- Probability P: With each event $A \in \mathcal{A}$, we associate a number $P(A)$ that measures the belief that the event will occur. Ex. $P\{\text{hh}, \text{tt}\} = 0.5$
- Target space $\mathcal{T}$: target quantities of interest. Ex. $\mathcal{T} = \{0, 1, 2\}$ possible heads.
- Random Variable $X : \Omega \to \mathcal{T}$ lets us convert probabilities on the sample space Ω to probabilities on targets $\mathcal{T}$ (i.e. $\mathcal{T} = \mathbb{R}$).
- If $S \subseteq \mathcal{T}$, then $P_X(S) = P(\{\omega \in \Omega : X(\omega) \in S\})$. P_X is the distribution of random variable X.
- If $\mathcal{T}$ is finite, X is a discrete random variable. If $\mathcal{T}$ is continuous, X is a continuous random variable.

Aside: Data points $x_1, \ldots, x_N$ are *observations* of a random variable (i.e. each observation is the result of an experiment). Probability lets us reason over these random experiments as $n \to \infty$. This will be key for studying generalization.

- Probability mass function: For a discrete random variable and a a potential observation $x \in \mathcal{T}$, we can write $P_X(x) = P(X = x)$. We often take X to be implicit and people just write $P(x)$.
- Joint probability: we can consider probabilities of multiple random variables, i.e. $P(X = x, Y = y) = P(X = x \cap Y = y)$, often abbreviated as $p(x, y)$.
- For example: for a dataset of examples with labels $(x_1, y_1), \ldots, (x_N, y_N)$ we can let X be a random variable for examples x_i, and Y be a random variable for the labels y_i. This lets us formalize the probability of observing an example and its label as $p(x_i, y_i)$.
- Marginal probability: the marginal of X is $P(X = x)$, which is irrespective of the random variable Y, often lazily written as $p(x)$
- Conditional probability: the conditional probability of Y given X is $P(Y = y|X = x)$, often lazily written as $p(y|x)$

Aside: In ML we will want to be modeling predictions from your data, i.e. $p(y|x)$ where x is the input to your model and y is the prediction. We'll now consider real valued continuous distributions, where $\mathcal{T} = \mathbb{R}^D$.

- Probability density function: A function $f : \mathbb{R}^D \to \mathbb{R}$ is a PDF if (1) $\forall x \in \mathbb{R}^D : f(x) \geq 0$ and (2) $\int_{\mathbb{R}^D} f(x)dx = 1$.
- This is like the probability mass function, where the integral is replaced by a sum.
- We can associate a PDF with a random variable X, in the 1D case, $P(a \leq X \leq b) = \int_a^b f(x)dx$ where $a, b, x \in \mathbb{R}$. In this case, P is the distribution of X.
- The multi-dimensional PDF is similar:

$$P(a_1 \leq X_1 \leq b_1, \ldots, a_D \leq X_D \leq b_D) = \int_{a_1}^{b_1} \cdots \int_{a_D}^{b_D} f(x_1, \ldots, x_D)dx_D \ldots dx_1$$

  We will often abbreviate these to vectors $a, b, x \in \mathbb{R}^D$ as $\int_a^b f(x)dx$
- $P(X = x)$ no longer makes sense here as it is equal to 0 (equivalent to taking the interval $[a, b]$ where $a = b = x$.).
- Typically we use a one-sided interval: for a particular outcome $x \in \mathcal{T}$, we often refer to to $F_X(x) = P(X \leq x)$ as the cumulative density function (CDF).
- For a vector of random variables (i.e. the joint distribution), this can be explicitly written out as $P(X_1 \leq x_1, \ldots, X_D \leq x_d)$, but will typically also be abbreviated as $F_X(x) = P(X \leq x)$.

Aside: all probabilities, discrete or continuous, are positive and sum to one. But for continuous distributions, the PDF may be more than one at some points. See Example 6.3 in the textbook on the uniform distribution.

- There are really only two fundamental rules in probability for reasoning about distributions: the sum and the product rule.
- The sum rule is also known as the marginalization property (recall the marginal of x is $p(x)$) and relates the joint distribution to the marginal distribution
- Sum rule (discrete):

$$p(x) = \sum_{y \in \mathcal{Y}} p(x, y)$$

- Sum rule (continuous):

$$p(x) = \int_{y \in \mathcal{Y}} p(x, y)dy$$

- Note that, as always, x, y can be vectors
- Example: for a joint distribution $p(x) = p(x_1, \ldots, x_D)$, we can sum out all but one to get $p(x_i) = \int p(x)dx_{\setminus i}$

- The product rule says that every joint distribution can be factorized into a product of a conditional and marginal.
- Product rule (discrete and continuous):

$$p(x, y) = p(y|x)p(x)$$

- Ordering of x, y is arbitrary, and uses PDF/PMF for continuous/discrete distributions
- Bayes theorem:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

- Consequence of the product rule
- In ML terms, this relates the posterior with the likelihood, prior and evidence (Eq 6.23 from the textbook)
- Prior $p(x)$: subjective belief about target of interest x without observing anything
- Likelihood $p(y|x)$ relates the evidence y and the target of interest x (likelihood of x given y)
- Posterior $p(x|y)$ is what we know about x after seeing the evidence y and is usually what we care about
- Evidence $p(y)$ keeps the distribution normalized, sometimes called the marginalized likelihood since $p(y) = \int p(y|x)p(x)dx$. This can be hard to compute for vector valued x.
- Bayes theorem lets use "invert" a conditional. This can be useful when the target we care about x is not directly observable, other evidence y is observable. By choosing a prior $p(x)$, we can reason about $p(x|y)$ in terms of only the evidence y without explicitly observing x.

Aside: A common application of this section of probability is to reason about the parameters of your hypothesis class and find the most likely set of parameters given the data (hence maximum likelihood). Recall that in ML, we try to select the "best" function from the hypothesis class $f_\theta \in \mathcal{F}$ where θ is a set of parameters that determines the exact function. Some hypothesis classes directly parameterize a probability.

For example, consider a linear model to predict one of two classes, i.e. $f(x) = p(1|x) = \sigma(\theta^T x)$ where $\sigma(x) = \frac{1}{1+e^{-x}}$. The sigmoid here just forces the output to be between 0 and 1, and therefore a valid probability. Our goal is to find θ, however, we do not observe directly what θ is. Instead we get some input/output examples $(x_1, y_1), \ldots, (x_N, y_N) = (X, Y)$, and therefore we want to find the best θ that fits the data. The Bayes theorem for this scenario is would typically be framed as:

$$p(\theta|X, Y) = \frac{P(X, Y|\theta)p(\theta)}{p(X, Y)}$$

Then, a learning algorithm will search for a θ that maximizes $p(\theta|X, Y)$.