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.

- Recall that in ML, we try to minimize the risk on a dataset:

$$\min_\theta R_{\text{emp}}(f_\theta, X, Y) = \sum_{i=1}^{N} \ell(f(x_i), y_i)$$

  where θ is a set of parameters that determines a particular function from the hypothesis class, $f_\theta \in \mathcal{F}$
- Often, in ML we define a function class that directly predicts $f_\theta(x) = p(y|x; \theta)$

- For example, a simple linear model that predicts one of two classes is

$$f(x) = p(1|x) = \sigma(\theta^T x)$$

  where σ is the sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

- The sigmoid function parameterizes a smooth function that transitions between $[0, 1]$
- One common choice for loss function is called the *negative log likelihood*: $\ell(f(x), y) = -\log p(y|x; \theta)$.
- Maximizing/minimizing the log of a function results in the same solution as maximizing/minimizing the original function
- This results in *maximum likelihood estimation* (MLE):

$$\min_\theta R_{\text{emp}}(f_\theta, X, Y) = \max_\theta \sum_{i=1}^{N} \log p(y|x; \theta)$$

- Example: MLE for the linear model is

$$\max_\theta \sum_{i=1}^{N} \log p(y|x; \theta) = \max_\theta \sum_{i=1}^{N} \log \sigma(\theta^T x_i) = \max_\theta \sum_{i=1}^{N} -\log(1 + e^{\theta^T x_i})$$

- In ML we call this (MLE + linear model + binary classification) logistic regression. Note that even though this is called logistic *regression*, it is confusingly predicting a *classification* problem.

In this example, MLE/negative log likelihood defines the objective, while linear binary classifier defines the model. This is just one possible example—many different machine learning models simply vary the the type of model. However, we can also vary the objective. A common choice is to use Bayes rule and is called maximum a posterior estimation (MAP).

- An alternative type of risk to minimize is the Bayes risk:

$$\min_\theta R_{\text{Bayes}}(f_\theta, X, Y) = \min_\theta -\log p(\theta|X, Y) = \max_\theta \log p(\theta|X, Y)$$

- This is in contrast to the empirical risk:

$$\min_\theta R_{\text{emp}}(f_\theta, X, Y) = \max_\theta \log p(Y|X, \theta)$$

- To calculate the Bayes risk, we simply apply Bayes rule to get something similar to the empirical risk:

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

- Minimizing the Bayes risk is known as *maximum a posterior estimation* (MAP):

$$\max_\theta \log p(Y|X; \theta) + \log p(\theta|X) - \log p(Y|X) \propto \max_\theta \log p(Y|X; \theta) + \log p(\theta)$$

- This differs from MLE only via the prior term $\log p(\theta)$
- The posterior in MAP in this case is $p(\theta|X, Y)$, hence maximum a posterior
- Whereas the MLE maximizes the likelihood, $p(Y|X, \theta)$

Aside: Up to this point, we've used random variables to represent examples from a given distribution, such as $(X_1, Y_1), \ldots, (X_N, Y_N)$ to represent N datapoints with N labels. However, we typically want to summarize these sets of random variables with a single quantity. This is called a *statistic*, which is a deterministic function of random variables. These statistics describe how random variables behave.

- Two common statistics: mean and variance
- Expected value of a function $g : \mathbb{R} \to \mathbb{R}$ of random variables is the average over many random draws. For continuous distributions this is:

$$\mathbb{E}_X[g(x)] = \int_{\mathcal{X}} g(x)p(x)dx$$

  For discrete distributions, this is:

$$\mathbb{E}_X[g(x)] = \sum_{\mathcal{X}} g(x)p(x)dx$$

- Sometimes, this is written as $\mathbb{E}_X[g(x)] = \mathbb{E}_{x\sim X}[g(x)] = \mathbb{E}[g(x)]$
- If X is a random variable with probability p, then we can also write this as $E_X[g(x)] = E_{p(x)}[g(x)]$ or $E_p[g(x)]$ or $E_{x\sim p}p[g(x)]$
- A conditional expectation is the same, using a conditional probability distribution:

$$\mathbb{E}_X[g(x)|y] = \int_{\mathcal{X}} g(x)p(x|y)dx$$

- An expectation of a vector of random variables is the vector of expectations of each random variables:

$$\mathbb{E}_X[g(x)] = \begin{bmatrix} \mathbb{E}_{X_1}[g(x_1)] \\ \vdots \\ \mathbb{E}_{X_N}[g(x_N)] \end{bmatrix}$$

- The mean statistic is the special case where $g(x) = x$, for example $\mathbb{E}[x] = \int_{\mathcal{X}} xp(x)dx$
- Often we use the symbol $\mathbb{E}[x] = \mu$
- Intuitively, the mean is the "average" value. We will use averages when summing many random variables together from the same distribution.
- The expected value is a *linear operator*. This means that if $f(x) = ag(x) + bh(x)$, then

$$\mathbb{E}[f(x)] = a\mathbb{E}[g(x)] + b\mathbb{E}[h(x)]$$

- Covariance is the expected product of deviations of two random variables from their means.

$$\text{Cov}_{X,Y}[x, y] = \mathbb{E}_{X,Y}[(x - \mathbb{E}_X[x])(y - \mathbb{E}_Y[y])$$

- Covariance measures how dependent two random variables are. If it is high, they more dependent

$$\text{Cov}_{X,Y}[x, y] = \mathbb{E}_{X,Y}[xy] - \mathbb{E}_X[x]\mathbb{E}_Y[y]$$

- The covariance of a variable with itself is the variance $\text{Var}_X[x] = \mathbb{V}[x] = \text{Cov}_{X,X}[x, x]$
- Often we use the symbol $\mathbb{V}[x] = \Sigma$
- For a single random variable, the square root of the variance is the standard deviation, $\sigma(x) = \sqrt{\text{Var}_X[x]}$
- Using the second form of the covariance, we can generalize this to vectors $x, y \in \mathbb{R}^D \times \mathbb{R}^E$ as

$$\text{Cov}_{X,Y}[x, y] = \mathbb{E}_{X,Y}[xy^T] - \mathbb{E}_X[x]\mathbb{E}_Y[y]^T \in \mathbb{R}^{D\times E}$$

  and the variance is

$$\mathbb{V}[x] = \text{Cov}[x, x]$$

  , also called the covariance matrix (measures spread)

- Correlation is a normalized form of covariance between two random variables (i.e. the covariance is divided by the variance of the two random variables and measures how closely two variables change together):

$$\text{corr}[x, y] = \frac{\text{Cov}[x, y]}{\sqrt{V[x]V[y]}}$$

- Variance can be done in three ways:
  1. $\mathbb{V}[x] = \mathbb{E}[(x - \mu)^2]$ measures spread of a random variable
  2. $\mathbb{V}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2$ is the "raw score formula" that can be done in one pass but is numerically unstable
  3. $\frac{1}{N}\sum_{ij}(x_i - x_j)^2 = 2\left[\frac{1}{N}\sum_i x_i^2 - \left(\frac{1}{N}\sum_i x_i\right)\right]$ is the sum of pairwise differences
- $\mathbb{E}[x + y] = \mathbb{E}[x] + \mathbb{E}[y]$
- $\mathbb{E}[x - y] = \mathbb{E}[x] - \mathbb{E}[y]$
- $\mathbb{V}[x + y] = \mathbb{V}[x] + \mathbb{V}[y] + \text{Cov}[x, y] + \text{Cov}[y, x]$
- $\mathbb{V}[x - y] = \mathbb{V}[x] + \mathbb{V}[y] - \text{Cov}[x, y] - \text{Cov}[y, x]$
- If $y = Ax + b$ where x, y are random variables, then

$$E[y] = \mathbb{E}[Ax + b] = A\mathbb{E}[x] + b = A\mu + b$$

  and

$$\mathbb{V}[y] = \mathbb{V}[Ax + b] = \mathbb{V}[Ax] = A\mathbb{V}[x]A^T = A\Sigma A^T$$

Aside: In practice, we don't typically have the true distributions of X, Y but instead have a finite number of observations of the random variables $(x_1, y_1), \ldots, (x_N, y_N)$. Therefore, we will often estimate the an expected value with these samples by replacing the expected value with a summation:

$$\mathbb{E}[g(x)] \approx \frac{1}{N} \sum_{i=1}^{N} g(x_i)$$

Therefore, the empirical mean and empirical covariance are simply

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

and

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

- Independence: Two random variable X, Y are statistically independent if and only if $p(x, y) = p(x)p(y)$
- This implies the following:
  1. $p(y|x) = p(y)$
  2. $p(x|y) = p(x)$
  3. $\mathbb{V}[x + y] = V[x] + V[y]$
  4. $\text{Cov}[x, y] = 0$
- The converse is not true, i.e. if $\mathbb{E}[x] = 0$ and $\mathbb{E}[x^3] = 0$ and let $y = x^2$, then $\text{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] = \mathbb{E}[x^3] = 0$ (they are dependent but not correlated)
- A standard assumption in ML is that random variables are *independent and identically distributed* (i.i.d.), typically for the random variables representing the observations in the dataset $(X_1, \ldots, X_N)$
- This means that each random variable has the same distribution p, and that each random variable is independent from each other
- We used this earlier when defining the empirical risk: $p(Y|X; \theta) = \prod_{i=1}^{N} p(y_i|x_i; \theta)$
- Conditional independence: X, Y are conditionally independent given Z if and only if $p(x, y|z) = p(x|z)p(y|z)$ for all $z \in \mathcal{Z}$
- Alternatively, $p(x|y, z = p(x|z)$. This can be seen by using the product rule on the LHS and comparing it to the definition of conditional independence.

Aside: In the previous linear example, recall that we wrote the empirical risk as a sum of losses over N examples:

$$\min_{\theta} R_{\text{emp}}(f_\theta, X, Y) = \sum_{i=1}^{N} -\log p(y_i|x_i; \theta)$$

where the loss was the negative log likelihood. Where did this loss come from? We actually cheated a little bit: this is only true if we assume that the random variables are i.i.d. Without any assumptions, our starting point is to actually maximize the joint likelihood of the entire dataset under the model parameterized by θ (maximum likelihood):

$$\max_{\theta} p(Y|X;\theta)$$

However, modeling an entire dataset jointly is complicated! To simplify this, we assume that the vectors of (x_i, y_i) are independent and identically distributed. Independence allows us to consider the product over N random variables:

$$\max_{\theta} \prod_{i=1}^{N} p_i(y_i|x_i;\theta)$$

and then identical allows us to use a single model to represent all data points:

$$\max_{\theta} \prod_{i=1}^{N} p(y_i|x_i;\theta)$$

Lastly, we take a log for numerical stability since the solution to the maximum is the same:

$$\max_{\theta} \sum_{i=1}^{N} \log p(y_i|x_i;\theta)$$

Flipping to minimizing the negative objective gets us the minimum likelihood formulation:

$$\min_{\theta} \sum_{i=1}^{N} -\log p(y_i|x_i;\theta)$$