CIS3333: Mathematics of Machine Learning Fall 2025

Lecture: Probability: Review

Date: September 3rd, 2025 *Author: Surbhi Goel*

Acknowledgements. These notes are based on the notes by Eric Wong from Fall 2024.

Disclaimer. These notes are intended to accompany Chapter 6 of the book Mathematics for Machine Learning by Deisenroth, Faisal, and Ong, and not as a substitute for the book.

1 Probability Basics

Probability Spaces and Random Variables

- 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. $\Omega = \{\text{hh, tt, ht, th}\}$
- Event space $\mathcal{A}$: the space of potential results (events). Ex. the power set 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.

Connection to ML: 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.

Discrete Probability Functions

- 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 $p(y|x)$

Connection to ML: 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.

Continuous Probability Functions 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 $F_X(x) = P(X \leq x)$.

Important Note: 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.

Fundamental Rules of Probability

- 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):

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

- Note that, as always, x, y can be vectors
- Example: For a joint distribution $p(x) = p(x_1, \ldots, x_D)$, we can find the marginal distribution of a single variable x_i by integrating over all other variables. This process is called *marginalization*. We can write this as:

$$p(x_i) = \int \cdots \int p(x_1, \ldots, x_D) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_D$$

  A shorthand for this operation used in the book is $\int p(x) dx_{\setminus i}$, where $dx_{\setminus i}$ indicates integrating over all variables in the vector x except for x_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

Independence

- 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. $p(x|y) = p(x)$
- A standard and foundational assumption in Machine Learning is that the data points in our dataset, $(x_1, y_1), \ldots, (x_N, y_N)$, are **independent and identically distributed (i.i.d.)**. This means each data point is drawn independently from the same underlying distribution, and the outcome of one draw has no influence on any other. This assumption is the key that makes working with datasets practical, as we'll see in the MLE section.

- 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.

Bayes' Theorem and Its Components

- Bayes theorem:
$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$
- This is a direct consequence of the product rule. Recall that the product rule can be written in two ways: $p(x, y) = p(x|y)p(y)$ and $p(x, y) = p(y|x)p(x)$. Setting these equal gives $p(x|y)p(y) = p(y|x)p(x)$, and rearranging gives Bayes' theorem.
- 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 the evidence y given the target x)
- 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.

An Analogy: The Doctor’s Diagnosis
To make these terms concrete, consider a medical diagnosis. The **evidence** y is a symptom, like a ‘cough‘. The **target** x is the underlying illness, like the ‘flu‘. In machine learning, x corresponds to our model parameters θ and y corresponds to our data.

- **Prior** $p(x)$: The doctor’s belief *before* seeing the patient. In February, the prior for ‘flu‘, $p(\text{flu})$, is high. This is our initial belief about the model parameters θ before seeing data.
- **Likelihood** $p(y|x)$: How well the illness explains the symptom. The likelihood of a ‘cough‘ given the ‘flu‘, $p(\text{cough}|\text{flu})$, is very high. This term asks, ”How well do our parameters θ explain our dataset?”
- **Posterior** $p(x|y)$: The final diagnosis. Given a ‘cough‘, what is the probability of the ‘flu‘? This is our updated, final belief about the best parameters θ after seeing the data.
- **Evidence** $p(y)$: The overall probability of the symptom. How common is a ‘cough‘ in general? This term is a normalizing constant and is often ignored when we just need to find the *best* x.

With this framework, we can now define our primary goal in probabilistic machine learning: to find the ”best” set of parameters θ for our model. The first and most direct way to do this is called Maximum Likelihood Estimation.

Maximum Likelihood Estimation (MLE) The **Maximum Likelihood Estimation (MLE)** principle is to pick the parameters θ that maximize the likelihood of observing the data, $p(\mathbf{Y}|\mathbf{X};\theta)$. This approach is like a ”data purist” doctor who only considers the likelihood. For example, if a ‘cough‘ is the symptom, they might find that the likelihood $p(\text{cough}|\text{Rare Lungworm})$ is 95%, while $p(\text{cough}|\text{Flu})$ is only 90%. Based only on this evidence, they would diagnose the rare lungworm. In ML, this is analogous to overfitting: we find parameters that perfectly explain the training data, even if it means choosing a model that is inherently implausible.

Applying MLE to a Dataset The likelihood of the entire dataset is the joint probability $p(\mathbf{Y}|\mathbf{X};\theta)$. Modeling this directly is intractable. To make this practical, we rely on the **i.i.d. assumption** we discussed earlier. This assumption allows us to simplify the objective from a complex joint likelihood to a more manageable product of individual likelihoods:

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

For numerical stability and mathematical convenience, we maximize the log-likelihood instead. This turns the product into a sum, which is equivalent to minimizing the negative log-likelihood loss:

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

This final expression is just the empirical risk, R_{emp}, where the loss function ℓ is the negative log-likelihood.

- **Example: Deriving the Logistic Regression Objective**. Let's see how this works for logistic regression.
  - First, we define our probabilistic model for a single data point. We use the sigmoid function $\sigma(z) = (1 + e^{-z})^{-1}$ and labels $y_i \in \{-1, 1\}$ to get a compact formula for the probability of the correct label:

    $$p(y_i|x_i; \theta) = \sigma(y_i \theta^T x_i)$$

  - Next, we plug this into the MLE objective we just derived to get the final optimization problem for logistic regression:

    $$\max_\theta \sum_{i=1}^N \log \sigma(y_i \theta^T x_i) = \max_\theta \sum_{i=1}^N -\log(1 + e^{-y_i \theta^T x_i})$$

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

In MLE, the negative log-likelihood defines the objective and the linear model defines the hypothesis class. This is just one possible combination. We can also vary the objective. An alternative, which uses the prior, is Maximum A Posteriori (MAP) estimation.

Maximum A Posteriori (MAP) Estimation The MAP principle incorporates the prior to avoid the pitfalls of MLE. It's like the "experienced doctor" who multiplies the likelihood with their prior knowledge. While the lungworm has a higher likelihood, its prior probability, $p(\text{Rare Lungworm})$, is nearly zero. The flu, however, has both a high likelihood and a high prior ($p(\text{Flu})$ is high in winter). The resulting posterior probability for the flu is therefore much higher. In ML, MAP estimation finds this same balance, using a prior to regularize the model and prevent overfitting. This is done by maximizing the posterior probability, $p(\theta|\mathbf{X}, \mathbf{Y}) \propto p(\mathbf{Y}|\mathbf{X}; \theta)p(\theta)$. However, this is only effective if the prior is well-chosen. A misaligned prior (e.g., a doctor who wrongly believes a rare disease is common) can bias the model and lead to worse results than MLE.

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

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

- This is in contrast to the empirical risk:

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

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

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

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

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

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