Probability for ML

1 Introduction to Discrete Probability

1.1 Basic Rules and Conditional Probability

Recall from lecture, a Probability Space consist of the following:

- A sample space Ω: The set of all possible outcomes ω of an experiment
- An event space: The set of all subsets A of the sample space Ω
- A probability function $P\colon \Omega \to [0, 1]$ that maps outcomes to values between 0 and 1 (inclusive on both ends)

Furthermore, we go over a few fundamentals:

- Union (OR): The $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ is the Principle of Inclusion-Exclusion for two events. The key intuition is that if you just add $P(A)$ and $P(B)$, you've double-counted the region where they overlap $P(A \cap B)$, so you must subtract it once.
- Conditional Probability (Given): $P(A|B)$ is the single most important concept for machine learning. It's the formal way of asking, "How does my belief about event A change after I observe event B?" The formula $\frac{P(A \cap B)}{P(B)}$ has a clear intuition: we are restricting our universe of possibilities to just the outcomes in B. The numerator, $P(A \cap B)$, is the part of A that still lives in this new, smaller universe. We divide by $P(B)$ to re-normalize the probabilities so that the total probability of our new universe B is 1.
- Independence: This is a special, simplifying assumption. It means B gives you no information about A. Knowing B happened doesn't change your belief about A at all ($P(A|B) = P(A)$). The formula $P(A \cap B) = P(A)P(B)$ is the computational check for this property. In ML, we love to assume independence (e.g., in Naive Bayes) because it lets us model complex joint probabilities just by multiplying simpler marginal probabilities.
- Intersection (AND): $P(A \cap B)$ represents the probability of A *and* B occurring.
- Complement (NOT): $P(\bar{A}) = 1 - P(A)$. One can think of this intuitively as simply the probability of the "opposite" of the event A. For example, If $\Omega = \{H, T\}$ and $A = H$ = landing heads, then $\bar{A} = T$ = landing tails.

This problem is meant for you to go over the basic rules of probability, as well as conditional probability again.

1. You roll two standard, fair 6-sided dice. Let Ω be the uniform probability space of 36 outcomes. Let A be the event that the sum of the dice is **even**. Let B be the event that at least one die shows a **6**.
   (a) Find $P(A)$.
   (b) Find $P(B)$.
   (c) Find $P(A \cap B)$ (the probability the sum is even AND there is at least one 6).
   (d) Find $P(A|B)$ (the probability the sum is even *given* that there is at least 6).
   (e) Are events A and B independent? Justify your answer.

1.2 Bayes' Rule and Law of Total Probability

From lecture:

- **Law of Total Probability (LTP):** This is a "divide and conquer" strategy. If you want to find the probability of a complex event B, you can break the world into a set of mutually exclusive and exhaustive "cases" $(A_1, A_2, \dots)$. The LTP says you can find $P(B)$ by calculating the probability of B in each case $(P(B|A_i))$ and then taking a weighted average of those probabilities, where the weights are the probabilities of each case happening $(P(A_i))$.
- **Bayes' Rule**: This rule is the machine that flips the conditional. In ML, we often have a model that can tell us the likelihood of our data given a hypothesis, $P(\text{Data} \mid \text{Hypothesis})$. But what we want is the reverse: the probability of our hypothesis given the data we saw, $P(\text{Hypothesis} \mid \text{Data})$. Bayes' Rule lets us do this:
  - Prior $P(A_i)$: Your belief about hypothesis A_i before you see any data.
  - Likelihood $P(B|A_i)$: The probability of observing data B if hypothesis A_i were true. This is what your model typically defines.
  - Evidence $P(B)$: The total probability of observing the data, calculated using the Law of Total Probability. It's a normalization constant.
  - Posterior $P(A_i|B)$: Your updated belief about hypothesis A_i after observing the data. This is the result of "learning".

1. A screening test for a disease is 90% accurate for people who have it (i.e., $P(\text{Pos} \mid \text{Disease}) = 0.9$) and 90% accurate for people who don't (i.e., $P(\text{ Neg} \mid \text{No Disease }) = 0.9$).

   Suppose the general prevalence of the disease in the population is 1% (i.e., $P(\text{Disease}) = 0.01$). A person is chosen at random, given the test, and it comes back positive. What is the probability they actually have the disease, $P(\text{Disease} \mid \text{Pos})$?

2 Discrete Random Variables

By definition, a Random Variable (RV) $X : \Omega \to \mathbb{R}$ maps outcomes to real numbers. For discrete random variables, they map outcomes to integers (e.g., $-1, 2, 3$) only, whereas for continuous random variables, they map outcomes to real numbers (e.g., $\sqrt{2}, \pi, e$).

2.1 Expectation and Linearity

With random variables, we can take expectations of them. What that means is we want to find the "center of mass" or "balancing point" of its PMF. It's the single number that best summarizes the "center" of a distribution. We write the expectation of a random variable X as $\mathbb{E}[X]$.

A notable property of expectations is that they are linear! That is, for constants a, b and random variables X, Y, $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$. The most critical part is that this holds regardless of whether X and Y are independent. This allows us to break down a very complex random variable (e.g., total error) into a sum of simple parts (error per data point) and analyze them individually, *even if they are correlated*.

1. Suppose $X \sim \text{Binomial}(n, p)$ and $Y \sim \text{Poisson}(\lambda)$. Find $\mathbb{E}[\pi X + eY]$.

2.2 Variance and Sum of RVs

We can also find the variance of a random variable, which we denote as $Var(X)$! This is the expected squared distance from the mean. It quantifies the "spread" or "risk" of a distribution. The $\mathbb{E}[X^2] - (\mathbb{E}[X])^2$ formula is purely for computational convenience. The definition $\mathbb{E}[(X - \mathbb{E}[X])^2]$ is the one to remember conceptually.

Properties of Variance include:

- $Var(aX + b) = a^2 Var(X)$. This is intuitive: Adding b shifts the entire distribution, but doesn't change its spread (so b disappears). Multiplying by a stretches the distribution by a factor of a. Since variance is in squared units (because of the squared term in its definition), the new spread is a^2 times the old one.
- Variance of a Sum: This is where independence becomes critical. $Var(X + Y) = Var(X) + Var(Y)$ only if X and Y are independent (or at least uncorrelated). This is because the full formula has a covariance term, as we'll see next.

1. Let X_1 be the outcome of a single fair 6-sided die roll.
   (a) Compute $\mathbb{E}[X_1]$ and $Var(X_1)$.
   (b) Let $S_n = X_1 + \cdots + X_n$ be the sum of n independent fair die rolls. Compute $\mathbb{E}[S_n]$ and $Var(S_n)$.

3 Multivariate Random Variables

This section bridges probability with linear algebra. We stop thinking of n features as n separate RVs and start thinking of them as one **random vector** $X \in \mathbb{R}^n$. Its properties are described by a mean vector μ and a covariance matrix Σ.

3.1 Joint and Marginal Distributions

A joint PMF $p_{X,Y}(x, y)$ (like a matrix) describes all probabilities.

- Marginal PMF: $p_X(x) = \sum_y p_{X,Y}(x, y)$ (summing over rows).
- Covariance: $Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$. If $Cov(X, Y) = 0$, they are uncorrelated.
- General Variance of Sum: $Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$.

1. Let the joint PMF of discrete RVs X and Y be given by the following table (matrix), where X is the row and Y is the column:

$p(x, y)$	$y = 0$	$y = 1$	$y = 2$
$x = 0$	0.1	0.1	0.0
$x = 1$	0.2	0.3	0.1
$x = 2$	0.0	0.1	0.1

   (a) Find the marginal PMFs $p_X(x)$ and $p_Y(y)$.

   (b) Find $\mathbb{E}[X]$ and $\mathbb{E}[Y]$.

   (c) Find $\mathbb{E}[XY]$ and $Cov(X, Y)$.

   (d) Are X and Y independent? Justify your answer.

3.2 Random Vectors and Covariance Matrices

In ML, a data point is a vector $X \in \mathbb{R}^d$ (e.g., d features). We model this as a "random vector" rather than d separate RVs.

We have also the notion of a mean vector denoted by μ. This is the "center of mass" of the d-dimensional probability cloud. It's an $\mathbb{R}^d$ vector where each element is the expectation of one feature: $\mu = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_d] \end{bmatrix}^\top$ as well as a covariance matrix, denoted by Σ. This is the $d \times d$ matrix that describes the shape and orientation of the data cloud. The diagonal elements, Σ_{ii}, are the variances $\Sigma_{ii} = Var(X_i)$. The off-diagonal elements, Σ_{ij}, are the covariances $\Sigma_{ij} = Cov(X_i, X_j)$. This matrix is symmetric ($\Sigma = \Sigma^\top$). If it's a diagonal matrix, it means all features are uncorrelated.

As seen earlier, linearity holds as well! Namely, $\mathbb{E}[AX + b] = A\mu + b$ and $Var(AX + b) = A\Sigma A^\top$. This is a powerful, direct application of linear algebra to probability.

1. Let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ be a random vector with mean $\mu = \mathbb{E}[X] = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ and covariance matrix $\Sigma = Var(X) = \begin{bmatrix} 4 & 1 \\ 1 & 2 \end{bmatrix}$.

   Let $A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \\ 1 & -1 \end{bmatrix}$ and $b = \begin{bmatrix} 5 \\ 0 \\ 0 \end{bmatrix}$. Let $Y = AX + b$.

   (a) Find the expectation vector $\mathbb{E}[Y]$.

   (b) Find the covariance matrix $Var(Y)$.

4 Continuous Random Variables

Recall from lecture that for any continuous value x, $P(X = x) = 0$. For that reason we can't use a PMF. To cope with this, we resort to the CDF (Cumulative Distribution Function): $F_X(x) = P(X \leq x)$. This is the fundamental object that works for all RVs (discrete and continuous). It's a function that "accumulates" probability from $-\infty$ up to x, and the PDF (Probability Density Function): $f_X(x) = F'_X(x)$. This is simply the derivative of the CDF. By the Fundamental Theorem of Calculus, this means $P(a \leq X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(x)dx$.

The key intuition to understand is that $f_X(x)$ is not a probability. It is a density. It can be greater than 1. The area under the PDF, $f_X(x)dx$, is the *infinitesimal probability mass at point* x.

Expectation (LOTUS): $\mathbb{E}[g(X)] = \int g(x)f_X(x)dx$. This is just the continuous version of the weighted average. You integrate the value $g(x)$ against its infinitesimal probability mass $f_X(x)dx$.

1. Let X be a continuous RV with PDF $f_X(x) = \begin{cases} cx^2 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$.

   (a) Find the constant c that makes $f_X(x)$ a valid PDF.

   (b) Find the CDF $F_X(x)$.

   (c) Find $\mathbb{E}[X]$.

   (d) Find $Var(X)$.

5 Multivariate Gaussians and Applications in ML Context

The Multivariate Gaussian (MVG): This is the one distribution that is completely defined by the linear algebra objects μ and Σ. Its PDF formula,

$$f(x) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu))$$

is literally a quadratic form in the exponent. Σ^{-1} (the precision matrix) defines the shape of the elliptical "data cloud".

MMSE (Best Estimator): $\mathbb{E}[Y|X=x]$. This is the best possible predictor for Y given $X=x$ to minimize squared error. In general, this is a complex, non-linear function $g(x)$.

LLSE (Best Linear Estimator): This is the best predictor $L(x) = aX + b$ that is restricted to being a line. It's what you solve for in Linear Regression. We find a and b by using multivariable calculus to minimize $\mathbb{E}[(Y-(aX+b))^2]$. The solution is the linear regression formula.

The Magic of Gaussians: A massive result in statistics is that if (X, Y) are jointly Gaussian, the (potentially complex) MMSE is the (simple) LLSE. The best possible predictor is just a straight line. This is why Linear Regression is not just a simple approximation; it is the mathematically optimal solution under the very common assumption of Gaussian-distributed data. This will be elaborated on more during the Classical ML lecture, so don't worry if you don't understand this part fully.

1. Suppose a 2D feature vector $Z = \begin{bmatrix} X \\ Y \end{bmatrix}$ is modeled as a Multivariate Gaussian, $Z \sim \mathcal{N}(\mu, \Sigma)$, with:

$$\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} 4 & 3 \\ 3 & 9 \end{bmatrix}$$

   (a) What are $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $Var(X)$, $Var(Y)$, and $Cov(X, Y)$?

   (b) Are X and Y independent? Why or why not?

   (c) What is the best linear estimator (LLSE) of Y given $X = x$? That is, find $L(x)$ that minimizes $\mathbb{E}[(Y - L(X))^2]$ where $L(X) = aX + b$.

   (d) What is the best overall estimator (MMSE) of Y given $X = x$? That is, find $\hat{Y}(x) = \mathbb{E}[Y|X = x]$.

6 What to Submit

Please submit a single .zip file that includes your written answers and completed notebook. Upload the .zip file to the corresponding homework assignment on Gradescope.

Contributors

- Patrick Mendoza