CIS3333: Mathematics of Machine Learning Fall 2025

Lecture: Probability: Fundamentals

Date: September 10th, 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.

Expected Value and Mean

We often want to summarize sets of random variables with a single quantity. This is called a *statistic*, which is a deterministic function of random variables. These statistics, like the mean and variance, describe how random variables behave and will be essential for characterizing the distributions we see later.

- 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:

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

- **Law of Total Expectation:** The expectation of a random variable is equal to the expectation of its conditional expectation with respect to any other random variable:

$$\mathbb{E}[x] = \mathbb{E}[\mathbb{E}[x|y]]$$

- For a random vector $\mathbf{x} = [x_1, \ldots, x_N]^T$, the expectation of a scalar function g applied element-wise is the vector of element-wise expectations:

$$\mathbb{E}[g(\mathbf{x})] = \begin{bmatrix} \mathbb{E}[g(x_1)] \\ \vdots \\ \mathbb{E}[g(x_N)] \end{bmatrix}$$

- The mean of a random vector $\mathbf{x}$ is a special case where $g(x) = x$. It is often denoted by $\boldsymbol{\mu}$:

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} = \int_{\mathcal{X}} \mathbf{x} p(\mathbf{x}) d\mathbf{x}$$

- 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 and Variance

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

$$\mathrm{Cov}[x, y] = \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])])$$

  This can be expanded to a more common computational form:

$$\mathrm{Cov}[x, y] = \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]$$

- Covariance measures the linear dependence between two random variables. A high value suggests a stronger linear relationship.

- However, two variables can be statistically dependent but have zero covariance if their relationship is non-linear. For example, let x be a random variable uniformly distributed on $[-1, 1]$, and let $y = x^2$. Clearly, y is dependent on x. However, their covariance is zero:

$$\begin{aligned} \mathrm{Cov}[x, y] &= \mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y] \\ &= \mathbb{E}[x^3] - \mathbb{E}[x]\mathbb{E}[x^2] \\ &= 0 - 0 \cdot \mathbb{E}[x^2] = 0 \end{aligned}$$

  since $\mathbb{E}[x] = 0$ and $\mathbb{E}[x^3] = 0$ for a uniform distribution symmetric around 0.

- The covariance of a variable with itself is the variance $\mathrm{Var}[x] = \mathbb{V}[x] = \mathrm{Cov}[x, x]$

- Often we use the symbol $\mathbb{V}[x] = \Sigma$

- For a single random variable, the square root of the variance is the standard deviation, $\sigma(x) = \sqrt{\mathrm{Var}[x]}$

- We can generalize this to vectors $\mathbf{x} \in \mathbb{R}^D$ and $\mathbf{y} \in \mathbb{R}^E$ as

  $$\text{Cov}[\mathbf{x}, \mathbf{y}] = \mathbb{E}[(\mathbf{x} - \mathbb{E}[\mathbf{x}])((\mathbf{y} - \mathbb{E}[\mathbf{y}])^T)] = \mathbb{E}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}]^T \in \mathbb{R}^{D \times E}$$

  and the variance is

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

  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

- For random vectors $\mathbf{x}, \mathbf{y}$:

- $\mathbb{E}[\mathbf{x} + \mathbf{y}] = \mathbb{E}[\mathbf{x}] + \mathbb{E}[\mathbf{y}]$

- $\mathbb{E}[\mathbf{x} - \mathbf{y}] = \mathbb{E}[\mathbf{x}] - \mathbb{E}[\mathbf{y}]$

- $\mathbb{V}[\mathbf{x} + \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}] + \text{Cov}[\mathbf{x}, \mathbf{y}] + \text{Cov}[\mathbf{y}, \mathbf{x}]$

- $\mathbb{V}[\mathbf{x} - \mathbf{y}] = \mathbb{V}[\mathbf{x}] + \mathbb{V}[\mathbf{y}] - \text{Cov}[\mathbf{x}, \mathbf{y}] - \text{Cov}[\mathbf{y}, \mathbf{x}]$

- If $\mathbf{y} = A\mathbf{x} + \mathbf{b}$ where $\mathbf{x}$ is a random vector, A is a matrix, and $\mathbf{b}$ is a vector, then $\mathbf{y}$ is a random vector, and

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

  and

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

Practical Implementation: 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:

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

Therefore, the empirical mean and empirical covariance are simply

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

and

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

For scalar random variables, another way to compute the empirical variance is via the sum of pairwise differences:

$$\frac{1}{N^2}\sum_{i,j}(x_i - x_j)^2 = 2\left[\frac{1}{N}\sum_i x_i^2 - \left(\frac{1}{N}\sum_i x_i\right)^2\right]$$

Gaussian/Normal Distribution

The Gaussian (or Normal) distribution is one of the most common and important in ML. First, the **Central Limit Theorem** states that the sum of many independent random variables tends toward a Gaussia (we will talk about this in more detail later), which is why many natural phenomena (e.g., height, measurement errors) are well-modeled by it. Second, for a fixed mean and variance, the Gaussian has the **maximum entropy**, making it the most "generic" or uninformative choice.

- For a single random variable (univariate case), the distribution is defined by its mean μ and variance σ^2:

  $$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}$$

  The term $(x-\mu)^2/\sigma^2$ measures the squared distance from the mean, scaled by the variance.

- For a D-dimensional random vector $\mathbf{x} \in \mathbb{R}^D$ (multivariate case), the distribution is defined by a mean vector $\boldsymbol{\mu} \in \mathbb{R}^D$ and a covariance matrix $\Sigma \in \mathbb{R}^{D\times D}$:

  $$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

  Here, $|\Sigma|$ is the determinant of the covariance matrix. The term in the exponent, $(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})$, is the squared *Mahalanobis distance*. It generalizes the univariate squared distance by accounting for the covariance between variables, measuring the distance from $\mathbf{x}$ to the mean $\boldsymbol{\mu}$ in a way that considers the shape of the distribution.

- We often use the shorthand notation $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

- The special case $\mathcal{N}(\mathbf{0}, I)$, where $\mathbf{0}$ is the zero vector and I is the identity matrix, is called the standard normal distribution.

Properties of Multivariate Gaussian

- Joint distribution of MVN. Suppose we represent a MVN as the concatenation of two vectors of MVN:

  $$p(\mathbf{x}, \mathbf{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix}\right)$$

  where $\Sigma_{xy} = \mathrm{Cov}[\mathbf{x}, \mathbf{y}]$ and Σ_{xx}, Σ_{yy} are the marginal variances of $\mathbf{x}$ and $\mathbf{y}$

- Then the marginals $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{y}) d\mathbf{y} = \mathcal{N}(\boldsymbol{\mu}_x, \Sigma_{xx})$ and $p(\mathbf{y}) = \int p(\mathbf{x}, \mathbf{y}) d\mathbf{x} = \mathcal{N}(\boldsymbol{\mu}_y, \Sigma_{yy})$ are Gaussian

- And the conditional distribution $p(\mathbf{x}|\mathbf{y})$ is also Gaussian

  $$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{x|y}, \Sigma_{x|y})$$

  where

  $$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \Sigma_{xy}\Sigma_{yy}^{-1}(\mathbf{y} - \boldsymbol{\mu}_y)$$

  and

  $$\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$$

- Products of Gaussians is Gaussian:

  $$\mathcal{N}(\mathbf{x}|\mathbf{a}, A)\mathcal{N}(\mathbf{x}|\mathbf{b}, B) = c'\mathcal{N}(\mathbf{x}|\mathbf{c}, C)$$

  where $C = (A^{-1} + B^{-1})^{-1}$, $\mathbf{c} = C(A^{-1}\mathbf{a} + B^{-1}\mathbf{b})$, and $c' = \mathcal{N}(\mathbf{a}|\mathbf{b}, A + B)$ (see 6.5.2) Note that in the definition of c', it is convenient to write it as the density of another Normal distribution even though c' is not random

- A weighted sum of Gaussian random vectors is also Gaussian:

  $$p(a\mathbf{x} + b\mathbf{y}) = \mathcal{N}(a\boldsymbol{\mu}_x + b\boldsymbol{\mu}_y, a^2\Sigma_x + b^2\Sigma_y)$$

- Sums of Gaussian random vectors is a special case of the weighted sum where $a = b = 1$:

  $$p(\mathbf{x} + \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y, \Sigma_x + \Sigma_y)$$

- **Mixture of Gaussians (Sum of Densities):** This is different from the sum of random variables. A mixture distribution describes a process where we first select a component distribution, then draw a sample from it. For a mixture of two univariate Gaussians (a Gaussian Mixture Model or GMM), the density is:

  $$p(x) = \alpha p_1(x) + (1 - \alpha)p_2(x), \quad \text{where } p_i(x) = \mathcal{N}(x|\mu_i, \sigma_i^2)$$

  The resulting distribution $p(x)$ is *not* Gaussian. Its mean and variance can be understood using the **law of total variance**. Let Z be a latent variable that selects component 1 with probability α and component 2 with probability $1 - \alpha$. Using the total law of expectation, the mean is the weighted average of the component means: $\mathbb{E}[x] = \alpha\mu_1 + (1 - \alpha)\mu_2$. The variance can be computed using the law of total variance:

  $$\mathbb{V}[x] = \mathbb{E}_Z[\mathbb{V}[x|Z]] + \mathbb{V}_Z[\mathbb{E}[x|Z]]$$

- Linear transform of a Gaussian is Gaussian. If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, and $\mathbf{y} = A\mathbf{x}$ is Gaussian where

  $$\mathbb{E}[\mathbf{y}] = \mathbb{E}[A\mathbf{x}] = A\mathbb{E}[\mathbf{x}] = A\boldsymbol{\mu}$$

  and

  $$\mathbb{V}[\mathbf{y}] = \mathbb{V}[A\mathbf{x}] = A\mathbb{V}[\mathbf{x}]A^T = A\Sigma A^T$$

  so $p(\mathbf{y}) = \mathcal{N}(A\boldsymbol{\mu}, A\Sigma A^T)$

Application: MAP Estimation and ℓ_2-Regularization Recall from the previous lecture that MAP estimation is equivalent to maximizing the log-likelihood plus the log-prior:

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \log p(\mathbf{Y}|\mathbf{X};\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

A common choice for the prior is a zero-mean Gaussian, $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{0}, \sigma^2 I)$, which expresses a belief that model weights should be small. The log of this prior simplifies to:

$$\log p(\boldsymbol{\theta}) = -\frac{1}{2\sigma^2}\|\boldsymbol{\theta}\|_2^2 + \text{constant}$$

Plugging this into the MAP objective yields the familiar objective for ℓ_2-regularization (or Ridge Regression):

$$\hat{\boldsymbol{\theta}}_{\text{MAP}} = \arg\max_{\boldsymbol{\theta}} \left(\sum_{i=1}^{N} \log p(y_i|x_i;\boldsymbol{\theta}) - \frac{1}{2\sigma^2}\|\boldsymbol{\theta}\|_2^2 \right)$$

Thus, placing a Gaussian prior on the model parameters is mathematically equivalent to adding an ℓ_2-regularization penalty to the MLE objective. The prior's variance, σ^2, is inversely proportional to the regularization strength.

Other Distributions

- Bernoulli distribution: for a random variable X with target state $x \in \{0,1\}$, $\text{Ber}(\mu)$ is defined as
$$p(x;\mu) = \mu^x(1-\mu)^{1-x}$$
where $\mathbb{E}[x] = \sum_x xp(x) = \mu$ and $\mathbb{V}[x] = \sum_x (x-\mu)^2 p(x) = (1-\mu)^2\mu + \mu^2(1-\mu) = \mu(1-\mu)$
- Bernoulli simulates flipping a coin with probability μ of being heads.
- This trick of using exponents for Boolean variables is often used in ML
- Binomial distribution: for a random variable X with target states $1,\ldots,N$, $\text{Bin}(N,\mu)$ is defined as
$$p(m;N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m}$$
where $\mathbb{E}[m] = N\mu$ and $\mathbb{V}[m] = N\mu(1-\mu)$
- Binomial simulates flipping a coin with bias μ for$N$ times and counting the number of heads

Conjugate Priors: In Bayesian inference, a prior distribution is **conjugate** to a likelihood function if the resulting posterior distribution belongs to the same family of distributions as the prior. This is computationally convenient because it provides a closed-form solution for the posterior. The Beta-Binomial pair is a classic example of this relationship.

- **Beta Distribution:** The conjugate prior for the Binomial likelihood is the Beta distribution, which models a probability $\mu \in [0,1]$. It is defined with parameters $\alpha, \beta > 0$, which can be thought of as pseudo-counts of successes and failures:
$$p(\mu;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}$$

- **Example 6.11 (Beta-Binomial Conjugacy, MML 6.6.1):** For a Binomial likelihood $p(x|\mu)$ and a Beta prior $p(\mu|\alpha, \beta)$, the posterior is also a Beta distribution. The update simply adds the observed counts (x successes, $N - x$ failures) to the prior's pseudo-counts:

$$p(\mu|x) \propto p(x|\mu)p(\mu) \propto \mu^x(1-\mu)^{N-x}\mu^{\alpha-1}(1-\mu)^{\beta-1} \propto \text{Beta}(\mu|x+\alpha, N-x+\beta)$$

Motivation: The convenient properties we've seen, such as the conjugacy of the Beta-Binomial pair, are not coincidences. They arise because these distributions belong to a larger class called the **exponential family**. This family provides a unified framework for many common distributions and is uniquely defined by a powerful property: it is the only family of distributions that can be summarized from data using a small, fixed number of values, known as **finite-dimensional sufficient statistics**.

- A distribution belongs to the **exponential family** if it can be written in the form:

$$p(\mathbf{x}|\boldsymbol{\theta}) = h(\mathbf{x}) \exp(\boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}) - A(\boldsymbol{\theta}))$$

- The key components are:
  - $\boldsymbol{\theta}$ are the natural parameters of the distribution.
  - $\boldsymbol{\phi}(\mathbf{x})$ is the **sufficient statistic**. A function is a sufficient statistic if it contains all the information from the data $\mathbf{x}$ needed to estimate the parameters $\boldsymbol{\theta}$. For example, for a Gaussian, the sufficient statistics are $\mathbf{x}$ and $\mathbf{x}\mathbf{x}^T$, meaning you can summarize an entire dataset with just these values without losing information about the mean and variance.
  - $A(\boldsymbol{\theta})$ is the log-partition function, which acts as a normalizer to ensure the distribution integrates to 1.

- Gaussian and Bernoulli are examples of exponential family distributions (see Examples 6.13 and 6.14).

- **Key Property: A Unified Theory of Conjugacy.** The most important property of the exponential family is that *every member has a conjugate prior.* This prior is also a member of the exponential family. This is the deep reason why we see conjugacy in so many different pairs of distributions (like Beta-Binomial or Gaussian-Gaussian)—they are all part of this same underlying mathematical structure. This provides a recipe for deriving the conjugate prior for any distribution in this family, unifying what would otherwise seem like a collection of convenient coincidences.