CIS3333: Mathematics of Machine Learning Fall 2025

Lecture: Probability

Date: September 8th, 2025 *Author: Surbhi Goel*

Acknowledgements. These notes are based on the notes by Eric Wong from Fall 2023.

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.

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. $p(x|y) = p(x)$
  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: $p(\mathbf{Y}|\mathbf{X}; \theta) = \prod_{i=1}^{N} p(y_i|x_i; \theta)$
- 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.

From Joint Likelihood to Empirical Risk In the linear model example from the MLE section, recall that we wrote the empirical risk as a sum of losses over N examples:

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

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):

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

However, modeling an entire dataset jointly is complicated! To simplify this, we assume that the data points (x_i, y_i) are independent and identically distributed (i.i.d.). The *independence* assumption allows us to factorize the joint likelihood into a product of individual likelihoods. At this stage, each data point could come from a different distribution p_i:

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

The *identically distributed* assumption then asserts that all these distributions are the same, i.e., $p_i = p$ for all i. This allows us to use a single model for 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:

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

Gaussian/Normal Distribution

- (Multivariate) Gaussian/Normal distribution is one of the most commonly used distributions in ML
- It represents having most samples clustered around the mean with the ability to have outliers
- Univariate Gaussian - $p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ for $x \in \mathbb{R}$
- Multivariate Gaussian - $p(x|\mu,\Sigma) = (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$ for $x \in \mathbb{R}^D$
- We write $p(x) = \mathcal{N}(x|\mu,\Sigma)$, $p(x) = \mathcal{N}(\mu,\Sigma)$, $p \sim \mathcal{N}(x|\mu,\Sigma)$ or $X \sim \mathcal{N}(\mu,\Sigma)$
- $\mathcal{N}(0, I)$ is the standard normal distribution where I is the identity matrix

Properties of Multivariate Gaussian

- Joint distribution of MVN. Suppose we represent a MVN as the concatenation of two vectors of MVN:
$$p(x, y) = \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right)$$
where $\Sigma_{xy} = \text{Cov}[x, y]$ and Σ_{xx}, Σ_{yy} are the marginal variances of x and y
- Then the marginals $p(x) = \int p(x, y)dy = \mathcal{N}(\mu_x, \sigma_{xx})$ and $p_y = \int p(x, y)dx = \mathcal{N}(\mu_y, \sigma_{yy})$ are Gaussian
- And the conditional distribution $p(x|y)$ is also Gaussian
$$p(x|y) = \mathcal{N}(\mu_{x|y}, \Sigma_{x|y})$$
where
$$\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$$
and
$$\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$$

Applications: Applications of the conditional Gaussian distribution are classic algorithms such as the Kalman filter (which does nothing but compute Gaussian conditions from joints) and Gaussian processes (assume that observations from a function are jointly Gaussian to get a Gaussian posterior over functions).

- Products of Gaussians is Gaussian:
$$\mathcal{N}(x|a, A)\mathcal{N}(x|b, B) = c'\mathcal{N}(x|c, C)$$
where
$$C = (A^{-1} + B^{-1})^{-1}$$
$$c = C(A^{-1}a + B^{-1}b)$$
$$c' = (2\pi)^{-D/2}|A + B|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(a - b)^T(A + B)^{-1}(a - b)\right) = \mathcal{N}(a|b, A + B)$$
- 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 variables is also Gaussian:
$$p(ax + by) = \mathcal{N}(a\mu_x + b\mu_y, a^2\Sigma_x + b^2\Sigma_y)$$
- Sums of Gaussians is a special case of the weighted sum where $a = b = 1$:
$$p(x + y) = \mathcal{N}(\mu_x + \mu_y, \Sigma_x + \Sigma_y)$$

- This is related to but different from the sum of the densities: if $p(x) = \alpha p_1(x) + (1-\alpha)p_2(x)$ where p_1, p_2 are Gaussian. In this case, p is not Gaussian. The mean is similar, i.e. $\mathbb{E}[x] = \alpha\mu_1 + (1-\alpha)\mu_2$, but the variance is different (Theorem 6.12):

$$\mathbb{V}[x] = [\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2] + \left([\alpha\mu_1^2 + (1-\alpha)\mu_2^t] - [\alpha\mu_1 + (1-\alpha)\mu_2]^2\right)$$

- This is an example of the law of total variance, i.e.

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

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

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

and

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

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

Motivation: The Gaussian is a distribution that has a lot of very nice properties. Many operations of Gaussians also return Gaussians. However, there are many things in this world are not Gaussian (i.e. even the simple mixture of Gaussians is not Gaussian). There is a generalization of Gaussians called the *exponential family* that has similarly nice properties but allows for a more expressive set of 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 probability μ N times and counting the number of heads
- Beta distribution: for a random variable μ with target states $[0, 1]$, $\text{Beta}(\alpha, \beta)$ for $\alpha, \beta > 0$ is defined as

$$p(\mu; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \mu^{\alpha-1}(1-\mu)^{\beta-1}$$

where $\mathbb{E}[\mu] = \frac{\alpha}{\alpha+\beta}$ and $\mathbb{V}[\mu] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

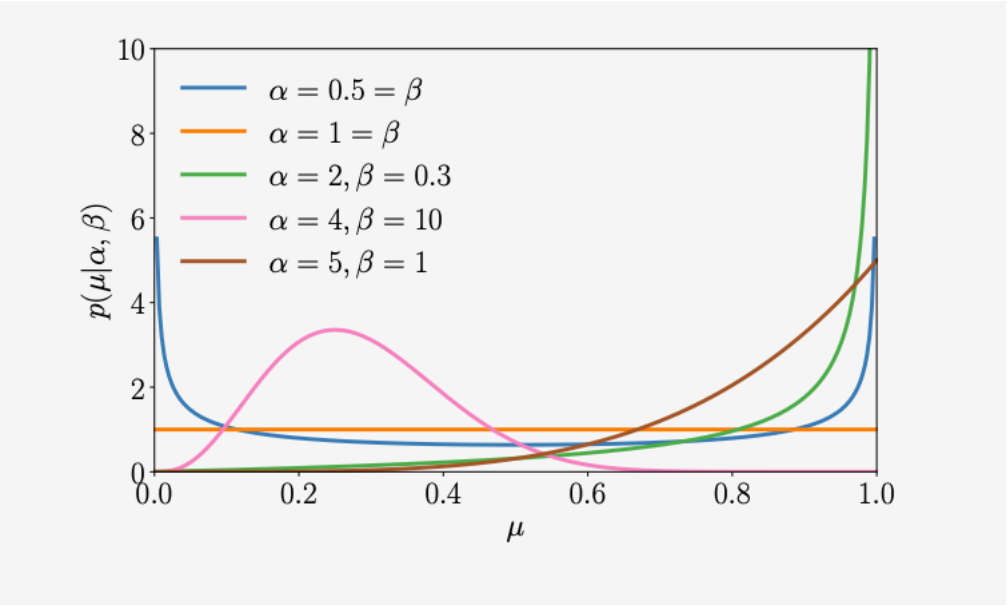

Figure 1: Beta distribution from Figure 6.2

- $\Gamma(t)$ is the Gamma function defined as

$$\Gamma(t) = \int_0^i nftyx^{t-1}\exp(-x)dx$$

  where $\Gamma(t+1) = t\Gamma(t)$.

- The Gamma function serves to normalize the Beta distribution.
- Beta models a continuous distribution on the interval $[0, 1]$ often used to simulate the probability of a binary evennt (i.e. the parameter of the Bernoulli distribution). Intuitively, $\alpha - 1$ can be thought of as the number of "successes" and $\beta - 1$ as the number of "failures" observed, shaping the distribution of the probability μ.
- $\alpha = \beta = 1$ is the Uniform distribution
- $\alpha, \beta < 1$ is bimodal with spikes at 0 and 1
- $\alpha, \beta > 1$ is unimodal. If $\alpha = \beta$ then it is symmetric and centered with mean/mode at 0.5.

Key Insight: We can define a whole slew of additional distributions. However there is a group of distributions that has a nice property like the Gaussian distribution, where combining two different distributions results in another known distribution. Remember that for the posterior, we have:

$$p(\theta|\mathbf{X}, \mathbf{Y}) \propto p(\mathbf{Y}|\theta, \mathbf{X})p(\theta)$$

If we say, model the likelihood as a Binomial and the prior as a Beta, it turns out that the posterior is a Beta distribution! This relation is known as conjugacy and shows up in the exponential family. Because the form of the posterior is nice and simple, ML algorithms like to use conjugate priors.

- Example 6.11 (Beta-Binomial Conjugacy)
- Suppose $x \sim \text{Bin}(N, \mu)$ (likelihood). Then consider a Beta prior on $\mu \sim \text{Beta}(\alpha, \beta)$. Then

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

- Similarly, Beta is also a conjugate prior for Bernoulli (Example 6.12). See Table 6.2 for more example of conjugate priors for common likelihoods.

Generalization: We can generalize these "nice" distributions to a larger family known as the exponential family. Exponential families interact nicely with the log operator (i.e. when calculating log probabilities) and have small tails (i.e. good for concentration around the mean). We'll abstract our distributions from parametric distributions with fixed parameters (i.e. $\mathcal{N}(0, 1)$) to parametric with learned parameters (i.e. $\mathcal{N}(\mu, \sigma^2)$ where μ, σ are estimated from the data with MLE) to families of distributions that capture multiple parametric forms (the exponential family).

- Sufficient statistic (Theorem 6.14, Fisher-Neyman): Let $X \sim p(x|\theta)$. Then, $\phi(x)$ is a sufficient statistic for θ if and only if $p(x|\theta)$ can be written as

$$p(x|\theta) = h(x)g_\theta(\phi(x))$$

  where $h(x)$ is independent of θ and g_θ captures all dependencies on θ via $\phi(x)$. In other words, $\phi(x)$ is a function of the data that captures all the information needed to estimate the parameter θ.

- Exponential family is characterized by

$$p(x|\theta) = h(x)\exp(\theta^T\phi(x) - A(\theta)) \propto \exp(\theta^T\phi(x))$$

  This is just a particular expression of g_θ for sufficient statistics.

- To see the last proportional equivalence, move $h(x)$ into the dot product by adding $\log h(x)$ to the sufficient statistics and add $\theta_0 = 1$ to the parameters, and $A(\theta)$ is just a normalizing constant, i.e.

$$A(\theta) = \log \int e^{\theta^T\phi(x)}dx$$

- Gaussian, Bernoulli are exponential families and have nice log probabilities. Example 6.13 (univariate Gaussian) and 6.14 (Bernoulli).
- Key property: every member of the exponential family has a conjugate prior (Brown, 1986):

$$p(\theta|\gamma) = h_c(\theta)\exp\left(\left\langle \begin{bmatrix} \gamma_1 \\ \gamma_2 \end{bmatrix}, \begin{bmatrix} \theta \\ -A(\theta) \end{bmatrix} \right\rangle - A_c(\gamma)\right)$$

- With this property, we can derive the conjugate prior without knowing it in advance for distributions in the exponential family