Probability for Deep Learning
Part 2

Johar M. Ashfaque

I. EXPECTATION

A. Expectation of a Random Variable

The expectation or the mean of a random variable X is the average value of X. The formal definition is as follows

Definition I.1 *The expected value or the mean or the first moment of X is defined to be*

$$\mathbb{E}(X) = \int x dF(x) = \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int_x x f(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

assuming that the sum (or integral) is well-defined. We use the following notation to denoted the expected value of X:

$$\mathbb{E}(X) = \mathbb{E}X = \int x dF(x) = \mu = \mu_X.$$

The expectation if a one-number summary of the distribution. From now on whenever we discuss expectations, we assume that they exist. Let $Y = r(X)$. How do we compute $\mathbb{E}(Y)$?

Theorem I.2 *Let $Y = r(X)$. Then*

$$\mathbb{E}(Y) = \mathbb{E}(r(X)) = \int r(x) dF_X(x).$$

This result makes intuitive sense. Think of playing a game where we draw X at random and then I pay you $Y = r(X)$. Your average income is $r(x)$ times the chance that $X = x$ summed or integrated over all values of x. The k^{th} moment of X is defined to be $\mathbb{E}(X^k)$ assuming that $\mathbb{E}(X^k) < \infty$. We shall rarely make much use of moments beyond $k = 2$.

B. Properties of Expectation

Theorem I.3 *If $X_1, ..., X_n$ are random variables and $a_1, ..., a_n$ are constants then*

$$\mathbb{E}\Big(\sum_i a_i X_i\Big) = \sum_i a_i \mathbb{E}(X_i).$$

Theorem I.4 *Let $X_1, ..., X_n$ be independent random variables. Then*

$$\mathbb{E}\Big(\prod_i^n X_i\Big) = \prod_i \mathbb{E}(X_i).$$

Notice that the summation rule does not require independence but the multiplication does.

C. Variance and Covariance

The variance measures the spread of the distribution.

Definition I.5 *Let X be a random variable with mean μ. The variance of X denoted σ^2 or σ_X^2 or $\mathbb{V}(X)$ is defined by*

$$\sigma^2 = \mathbb{E}(X-\mu)^2 = \int (x-\mu)^2 dF(x)$$

assuming this expectation exists. The standard deviation is $sd(X) = \sigma$.

Theorem I.6 *Assuming the variance is well-defined, it has the following properties:*

1. $\mathbb{V}(X) = \mathbb{E}(X^2) - E(X)^2 = \mathbb{E}(X^2) - \mu^2$.
2. *If a and b are constants then $\mathbb{V}(aX+b) = a^2\mathbb{V}(X)$.*
3. *If $X_1, ..., X_n$ are independent and $a_1, ..., a_n$ are constants then*

$$\mathbb{V}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i^2 \mathbb{V}(X_i).$$

If $X_1, ..., X_n$ are random variables then we define the sample mean to be

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

and the sample variance to be

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2.$$

Theorem I.7 *Let $X_1, ..., X_n$ be IID and let $\mu = \mathbb{E}(X_i)$, $\sigma^2 = \mathbb{V}(X_i)$. Then*

$$\mathbb{E}(\overline{X}_n) = \mu, \qquad \mathbb{V}(\overline{X}_n) = \frac{\sigma^2}{n}, \qquad \mathbb{E}(S_n^2) = \sigma^2.$$

If X and Y are random variables then the covariance and correlation between X and Y measure how strong the linear relationship is between X and Y.

Definition I.8 *Let X and Y be random variables with means μ_X and μ_Y and standard deviations σ_X and σ_Y. Define the covariance between X and Y by*

$$Cov(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$$

and the correlation by

$$\rho = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

Theorem I.9 *The covariance satisfies*

$$Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

The correlation satisfies

$$-1 \leq \rho(X,Y) \leq 1.$$

If $Y = a + bX$ for some constants a and b then $\rho(X,Y) = 1$ if $b > 0$ and $\rho(X,Y) = -1$ if $b < 0$. If X and Y are independent, then $\text{Cov}(X,Y) = \rho(X,Y) = 0$. The converse is not true in general.

Distribution	Mean	Variance
Point mass at a	a	0
Bernoulli (p)	p	$p(1-p)$
Binomial (n,p)	p	$np(1-p)$
Geometric (p)	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson (λ)	λ	λ
Uniform (a,b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal (μ,σ^2)	μ	σ^2
Exponential (β)	β	β^2
Gamma (α,β)	$\alpha\beta$	$\alpha\beta^2$
Beta (α,β)	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
t_ν	0 $(\nu>1)$	$\frac{\nu}{\nu-2}$ $(\nu>2)$
χ^2_p	p	$2p$