Introduction to Machine Learning

Lecture 3: Random Variables (Cont.)

Prof. Subrahmanya Swamy Peruru
Department of Electrical Engineering
IIT Kanpur

Recap

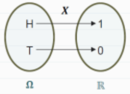
Random variable

- A random variable **maps** the **outcomes** of random experiment **to real numbers**

$X: \Omega \to \mathbb{R}$

- We often write $X(\omega)$ as the value of the RV when outcome ω occurs.

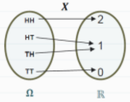
$X(H) = 1$
$X(T) = 0$



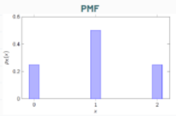
Example:

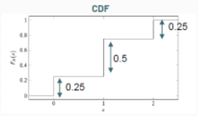
Experiment: Toss a coin twice
Random Variable: Number of Heads

$\Omega = \{HH, HT, TH, HH\}$



Discrete RVs: PMF and CDF





Discrete CDF is a **step function**

Jumps equal the PMF masses

$$\mathbb{P}(X \le x) = F_X(x) = \sum_{y=-\infty}^{x} p_X(y)$$

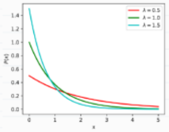
Examples : $\mathbb{E}[X]$ and $\mathrm{Var}(X)$

Uniform RV $X \sim Unif(0,1)$

- $\mathrm{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x) dx = \int_0^1 x\ dx = \frac{1}{2}$
- $\mathrm{E}[X^2] = \int_0^1 x^2\ dx = \frac{1}{3}$
- $\mathrm{Var}(X) = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$

Exercise: Solve this problem for $X \sim Unif(a,b)$

Exponential RV $X \sim \mathrm{Exp}(\lambda)$



- $f_X(x) = \lambda e^{-\lambda x}, \quad x \ge 0$
- $\mathrm{E}[X] = \int_0^{\infty} x\, \lambda e^{-\lambda x}\, dx = \frac{1}{\lambda}$
- $Var(X) = \frac{1}{\lambda^2}$

Multiple RVs: Joint PMF (discrete)

Suppose I want to characterize two random variables (X, Y)

Is it sufficient to know their PMFS: $p_X(x), p_Y(y)$?

NO

Example: Predict rain or not based on
$X = Temperature$
$Y = Humidity$

For discrete X, Y, **joint PMF**:

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y).$$

$$\sum_{x,y} p_{XY}(x, y) = 1$$

Example:

$X = Cointoss\ \{0, 1\},$
$Y = 3 - Faced\ die\ \{1, 2, 3\}$

	$y = 0$	$y = 1$	$y = 2$
$x = 0$	$p_{XY}(0,0)$	$p_{XY}(0,1)$	$p_{XY}(0,2)$
$x = 1$	$p_{XY}(1,0)$	$p_{XY}(1,1)$	$p_{XY}(1,2)$

Covariance and correlation

Covariance:

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

Correlation Coefficient:

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \in [-1,1].$$

Several sets of (x, y) points, with the correlation coefficient of x and y for each set.

The correlation captures
- the presence of a **linear relationship**
- and how **noisy** that relationship is

Uncorrelated RVs: X and Y are uncorrelated if $Cov(X, Y) = 0$ (or equivalently) $\rho_{X,Y} = 0$

Figure Source: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Lecture Content

1 Properties of Mean, Variance

- Properties of variance
- Law of Total Expectation
- Law of Total Variance

2 Random Vectors

- Covariance Matrix
- Linear Transformations
- Multivariate Gaussian

3 Entropy

- Entropy as a measure of uncertainty
- Cross entropy

Some properties of variance

Let X, Y have finite second moments and $a \in \mathbb{R}$.

1. Non-negativity

$$\mathrm{Var}(X) \geq 0, \qquad \mathrm{Var}(X) = 0 \Leftrightarrow X \text{ is almost surely constant.}$$

2. Scaling

$$\mathrm{Var}(aX + \mathrm{b}) = a^2 \mathrm{Var}(X)$$

3. Variance of a sum

$$\mathrm{Var}(aX + bY) = a^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(Y) + 2\, ab\, \mathrm{Cov}(X, Y)$$

Special case (independent / uncorrelated):

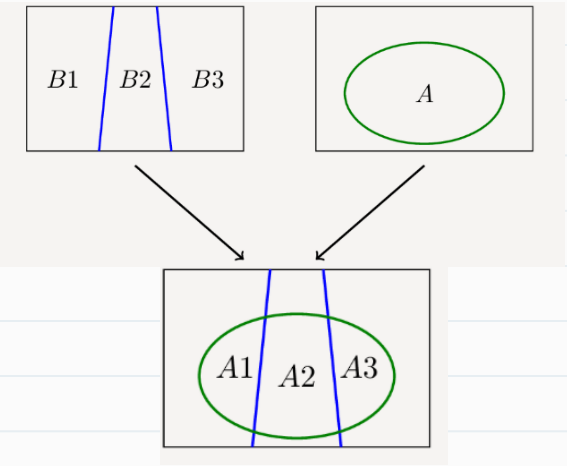
$$\mathrm{Cov}(X, Y) = 0 \quad \Rightarrow \quad \mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

In general, uncorrelated does not imply independent.

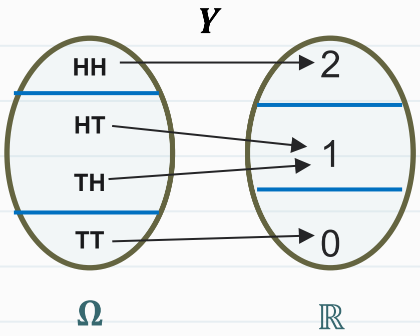
For Gaussian random variables, uncorrelated implies independent.

Law of total probability (RV)

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + P(A \mid B_3)P(B_3)$$



$$\mathbb{P}(\mathbf{X} = \mathbf{x})$$



$$\mathbb{P}(X = \mathrm{x}) = \sum_i \mathbb{P}(X = \mathrm{x} \mid Y = y_i)\, \mathbb{P}(Y = y_i).$$

Law of Total Expectation: An Example

Marks of 5 students: $10, 20, 30, 40, 50$.

Section split:

Section 1 has 2 students: marks 10,20

Section 2 has 3 students: marks 30,40,50

Check (direct average):

$$\frac{10 + 20 + 30 + 40 + 50}{5} = \frac{150}{5} = 30.$$

Compute section-wise averages:

$$\text{Avg}_{\text{Sec 1}} = \frac{10 + 20}{2} = 15, \qquad \text{Avg}_{\text{Sec 2}} = \frac{30 + 40 + 50}{3} = 40.$$

Overall class average as a weighted average:

$$\text{Overall Avg} = \left(\frac{2}{5}\right)\text{Avg}_{\text{Sec 1}} + \left(\frac{3}{5}\right)\text{Avg}_{\text{Sec 2}}$$

$$= \left(\frac{2}{5}\right) \cdot 15 + \left(\frac{3}{5}\right) \cdot 40 = 6 + 24 = 30.$$

Takeaway:

overall average = *weighted average of group averages*

(weights = fraction of students in each group).

A Probability Setup

Experiment: Pick *one* of the 5 students uniformly at random.

$$\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5\}, \qquad \mathbb{P}(\omega_i) = 1/5\,.$$

Random variables

$$Y(\omega) = \begin{cases} 1, & \omega \in \{\omega_1, \omega_2\} \quad \text{(Section 1)} \\ 2, & \omega \in \{\omega_3, \omega_4, \omega_5\} \quad \text{(Section 2)} \end{cases} \qquad X(\omega) \in \{10,20,30,40,50\}.$$

Goal: compute $\mathbb{E}[X]$

(A) Direct computation

$$\mathbb{E}[X] = \sum_{i=1}^{5} X(\omega_i)\,\mathbb{P}(\omega_i) = \frac{10 + 20 + 30 + 40 + 50}{5} = 30.$$

Law of Total Expectation

(B) Via conditioning on the section Y

First compute the conditional means:

$$\mathbb{E}[X \mid Y = 1] = \frac{10 + 20}{2} = 15, \qquad \mathbb{E}[X \mid Y = 2] = \frac{30 + 40 + 50}{3} = 40.$$

Also,

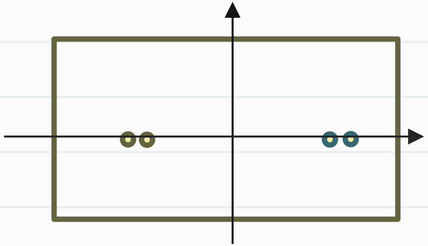
$$\mathbb{P}(Y = 1) = \frac{2}{5}, \qquad \mathbb{P}(Y = 2) = \frac{3}{5}.$$

Now apply **Law of Total Expectation**:

$$\boxed{\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X \mid Y]]}$$

$$\mathbb{E}[\mathbb{E}[X \mid Y]] = 15 \cdot \frac{2}{5} + 40 \cdot \frac{3}{5} = 30.$$

Law of Total Variance



$$\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y])$$

Interpretation: total spread = (average spread within sections) + (spread of section averages).

$$X \in \{-102, -100, 100, 102\},$$

Step 1: Overall variance

$$\mathbb{E}[X] = 0, \qquad \mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \frac{10000 + 10404 + 10000 + 10404}{4} = 10202$$

Law of Total Variance: Numerical Illustration

Random variable:

$$X \in \{-102,\ -100,\ 100,\ 102\}, \quad \mathbb{P}(X = x) = 1/4$$

Auxiliary RV Y:

$$Y = \begin{cases} 1, & X \in \{-102, -100\} \\ 2, & X \in \{100, 102\} \end{cases} \qquad \mathbb{P}(Y = 1) = \mathbb{P}(Y = 2) = 1/2$$

Step 2: Conditional variances

$$\mathbb{E}[X \mid Y = 1] = -101, \quad \mathrm{Var}(X \mid Y = 1) = 1$$

$$\mathbb{E}[X \mid Y = 2] = 101, \quad \mathrm{Var}(X \mid Y = 2) = 1$$

$$\mathbb{E}[\mathrm{Var}(X \mid Y)] = 1$$

Step 3: Variance of conditional mean

$$\mathrm{Var}(\mathbb{E}[X \mid Y]) = \mathrm{Var}(\{-101, 101\}) = 101^2 = 10201$$

$$\boxed{\begin{gathered}\mathrm{Var}(X) = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}(\mathbb{E}[X \mid Y]) \\ = 1 + 10201 = 10202\end{gathered}}$$

Example

Given:

$$X \sim \mathrm{Uniform}(1,2).$$

Conditioned on $X = x$, the RV Y is exponential with parameter $\lambda = x$:

$$Y \mid X = x \sim \mathrm{Exponential}(x).$$

Equivalently, we write:

$$Y \mid X \sim \mathrm{Exponential}(X).$$

Tasks:

1. Find $\mathbb{E}[Y]$.
2. Find $\mathrm{Var}(Y)$.

Example: Use law of total expectation

For an exponential RV with rate (parameter) λ:

$$\mathbb{E}[Y] = \frac{1}{\lambda}, \qquad \mathrm{Var}(Y) = \frac{1}{\lambda^2}.$$

So, conditioned on $X = x$:

$$\mathbb{E}[Y \mid X = x] = \frac{1}{x}, \qquad \mathrm{Var}(Y \mid X = x) = \frac{1}{x^2}.$$

Hence, as random variables (functions of X):

$$\mathbb{E}[Y \mid X] = \frac{1}{X}, \qquad \mathrm{Var}(Y \mid X) = \frac{1}{X^2}.$$

By the tower property (law of total expectation),

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}\left[\frac{1}{X}\right].$$

Since $X \sim \mathrm{Uniform}(1,2)$, its pdf is $f_X(x) = 1$ on $[1,2]$, so:

$$\mathbb{E}\left[\frac{1}{X}\right] = \int_1^2 \frac{1}{x}\, dx.$$

$$\mathbb{E}[Y] = \int_1^2 \frac{1}{x}\, dx = [\ln x]_1^2 = \ln 2.$$

$$\boxed{\mathbb{E}[Y] = \ln 2.}$$

Example: Use law of total variance

Law of total variance:

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X]).$$

We already have:

$$\mathrm{Var}(Y \mid X) = \frac{1}{X^2}, \qquad \mathbb{E}[Y \mid X] = \frac{1}{X}.$$

So we need:

$$\mathbb{E}\left[\frac{1}{X^2}\right] \quad \text{and} \quad \mathrm{Var}\left(\frac{1}{X}\right).$$

First,

$$\mathbb{E}\left[\frac{1}{X^2}\right] = \int_1^2 \frac{1}{x^2}\,dx = [-\frac{1}{x}]_1^2 = 1 - \frac{1}{2} = \frac{1}{2}.$$

Next,

$$\mathrm{Var}\left(\frac{1}{X}\right) = \mathbb{E}\left[\frac{1}{X^2}\right] - (\mathbb{E}\left[\frac{1}{X}\right])^2 = \frac{1}{2} - (\ln 2)^2.$$

Now plug into total variance:

$$\mathrm{Var}(Y) = \mathbb{E}\left[\frac{1}{X^2}\right] + \mathrm{Var}\left(\frac{1}{X}\right) = \frac{1}{2} + \left(\frac{1}{2} - (\ln 2)^2\right).$$

So:

$$\boxed{\mathrm{Var}(Y) = 1 - (\ln 2)^2.}$$

Random vector and mean vector

Random vector. Let

$$\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \qquad \text{(each } X_i \text{ is a random variable).}$$

Mean / expectation vector.

$$\boxed{\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}}$$

Linearity (component-wise). For fixed matrices/vectors and random vectors $\mathbf{X}, \mathbf{Z}$:

$$\boxed{\mathbb{E}[\mathbf{X} + \mathbf{Z}] = \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{Z}]}$$

$$\boxed{\mathbb{E}[A\mathbf{X} + b] = A\,\mathbb{E}[\mathbf{X}] + b}$$

Correlation matrix

Definition (Correlation matrix / second-moment matrix).

$$\boxed{R_{\mathbf{X}} \; = \; \mathbb{E}[\mathbf{X}\mathbf{X}^T]}$$

Entries:

$$R_{\mathbf{X}}(i,j) = \mathbb{E}[X_i X_j] \quad \text{(a cross moment).}$$

Example for $n = 3$:

$$R_{\mathbf{X}} = \begin{bmatrix} \mathbb{E}[X_1^2] & \mathbb{E}[X_1 X_2] & \mathbb{E}[X_1 X_3] \\ \mathbb{E}[X_2 X_1] & \mathbb{E}[X_2^2] & \mathbb{E}[X_2 X_3] \\ \mathbb{E}[X_3 X_1] & \mathbb{E}[X_3 X_2] & \mathbb{E}[X_3^2] \end{bmatrix}.$$

Covariance matrix

Definition (Covariance matrix).

$$\boxed{C_{\mathbf{X}} = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T]}$$

Entry-wise:

$$C_{\mathbf{X}}(i,j) = \mathrm{Cov}(X_i, X_j), \qquad C_{\mathbf{X}}(i,i) = \mathrm{Var}(X_i).$$

Key identity:

$$\boxed{C_{\mathbf{X}} = R_{\mathbf{X}} - \mathbb{E}[\mathbf{X}]\,\mathbb{E}[\mathbf{X}]^T}$$

where $\mu = \mathbb{E}[\mathbf{X}]$.

Linear transforms

Covariance matrices are always PSD.

For any fixed $b \in \mathbb{R}^n$,

$$\boxed{b^T C_{\mathbf{X}} b \;=\; \mathrm{Var}(b^T \mathbf{X}) \;\geq\; 0}$$

So $C_{\mathbf{X}}$ is symmetric and positive semi-definite.

Linear transform rule. Let $\mathbf{Y} = A\mathbf{X} + b$ where A and b are fixed. Then

$$\boxed{\mathbb{E}[\mathbf{Y}] = A\,\mathbb{E}[\mathbf{X}] + b}$$

$$\boxed{C_{\mathbf{Y}} = A\,C_{\mathbf{X}}\,A^T}$$

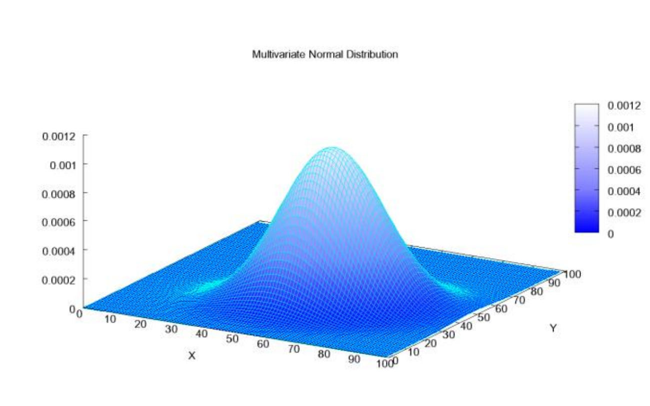
Intuition:
b shifts the mean but does not change “spread”;
A reshapes/rotates the spread.

Gaussian random vector: pdf

A random vector $\mathbf{X} \in \mathbb{R}^n$ is **Gaussian** (multivariate normal) with mean $\boldsymbol{\mu}$ and covariance Σ if

$$\boxed{f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}$$

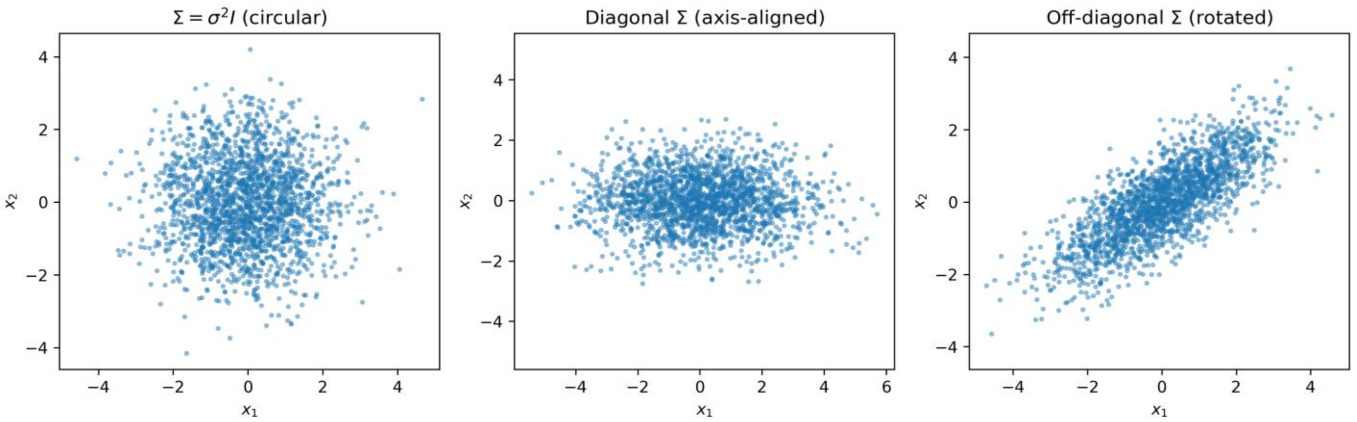
$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{X}]$$
$$\Sigma = \mathrm{Cov}(\mathbf{X})$$



Key fact: Gaussian is completely characterized by $(\boldsymbol{\mu}, \Sigma)$.

Figure Source: https://en.wikipedia.org/

Covariance controls contour shape

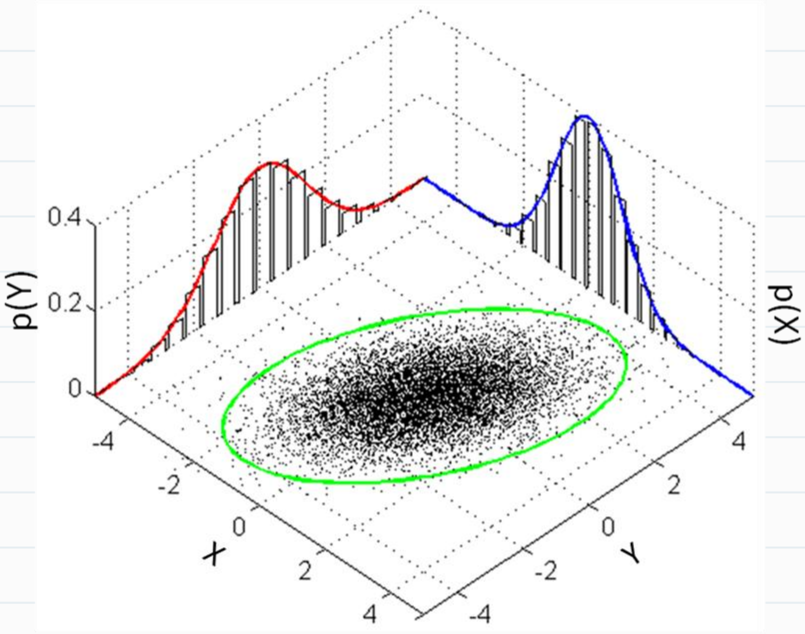


$$\Sigma = \begin{bmatrix} \mathbf{1.5} & 0 \\ 0 & \mathbf{1.5} \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \mathbf{3} & 0 \\ 0 & \mathbf{0.8} \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{2} & \mathbf{1.2} \\ \mathbf{1.2} & \mathbf{1.2} \end{bmatrix}$$

Marginal distribution is also Gaussian



**A multivariate normal distribution,
along with the two marginal distributions**

Figure Source: https://en.wikipedia.org/

Entropy

Definition (Discrete RV). For a discrete random variable X with pmf $p(x)$,

$$\boxed{H(X) = -\sum_{x} p\,(x) \log p(x)}$$

Interpretation: average uncertainty / surprise in outcomes.

Coin examples (base-2 logs ⇒ bits):

Fair coin: $p(H) = 0.5 \Rightarrow H(X) = 1$ bit (maximum uncertainty)

Biased coin: $p(H) = 0.9 \Rightarrow H(X) < 1$ bit (more predictable)

Takeaway: more randomness ⇒ higher entropy.

Cross entropy: comparing two distributions

Definition. Let $p(x)$ be the (true) data distribution and $q(x)$ a model distribution:

$$\boxed{H(p,q) = -\sum_x p(x) \log q(x)}$$

Interpretation:

Smaller $H(p,q)$ means q matches p better

ML insight: minimizing cross-entropy ≈ making model output q close to data distribution p.

Cross entropy: Numerical example

Numerical example.

Let

$$p = [0.8,\ 0.2], \qquad q_1 = [0.75,\ 0.25], \qquad q_2 = [0.4,\ 0.6].$$

Cross entropy with q_1:

$$\begin{aligned} H(p, q_1) &= -(0.8 \log(0.75) + 0.2 \log(0.25)) \\ &\approx -(0.8(-0.288) + 0.2(-1.386)) \\ &\approx 0.51 \end{aligned}$$

Cross entropy with q_2:

$$H(p, q_2) \approx 0.84$$

$$\boxed{H(p, q_1) < H(p, q_2)}$$

Conclusion: q_1 matches p better than q_2 ,hence it has lower cross entropy.

Some common discrete distributions

Bernoulli(p) (single yes/no trial)

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p \qquad \mathbb{E}[X] = p, \quad \mathrm{Var}(X) = p(1 - p)$$

Binomial(n, p) (# successes in n trials)

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k},\ k = 0,1, \dots, n \qquad \mathbb{E}[X] = np,\ \mathrm{Var}(X) = np(1 - p)$$

Geometric(p) (trials until first success)

$$\mathbb{P}(X = k) = (1 - p)^{k-1} p,\ k = 1,2, \dots \qquad \mathbb{E}[X] = \frac{1}{p},\ \mathrm{Var}(X) = \frac{1 - p}{p^2}$$

Poisson(λ) (event counts in a fixed interval)

$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!},\ k = 0,1,2, \dots \qquad \mathbb{E}[X] = \lambda,\ \mathrm{Var}(X) = \lambda$$

Further Reading: Categorical, Multinomial

Common continuous random variables

Uniform(a, b) (no preference in $[a, b]$)

$$f_X(x) = \frac{1}{b-a}\,\mathbf{1}\{a \leq x \leq b\} \qquad \mathbb{E}[X] = \frac{a+b}{2},\ \mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

Gaussian(μ, σ^2) (noise / errors; CLT)

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \mathbb{E}[X] = \mu,\ \mathrm{Var}(X) = \sigma^2$$

Exponential(λ) (time to event; memoryless)

$$f_X(x) = \lambda e^{-\lambda x}\,\mathbf{1}\{x \geq 0\} \qquad \mathbb{E}[X] = \frac{1}{\lambda},\ \mathrm{Var}(X) = \frac{1}{\lambda^2}$$

Further Reading: Laplace, Beta distribution, Gamma Distribution

References

- *Chapters 3-6* of the book **H. Pishro-Nik, "Introduction to probability, statistics, and random processes",** available at https://www.probabilitycourse.com, Kappa Research LLC, 2014.

Thank You