Machine Learning

Probability Review
for Discrete Random Variables

Fabio Vandin October 7^{th}, 2022

Expected Value and Moments

Definition

The expectation of a discrete random variable X is
$\mathbb{E}[X] = \sum_x x p_X(x)$.

r.v. r.v value

Let X be a r.v. Then $\mathbb{E}[X] \in \mathcal{O}$, where
$\mathcal{O} = \{$values taken by $X\}$.

TRUE? 3

FALSE? 72

Example

1) die rolling: X = outcome of a die

$$\mathbb{E}[X] = \sum_{i=1}^{6}\left(i \Pr[X=i]\right) = \sum_{i=1}^{6}\left(i \cdot \frac{1}{6}\right) = \frac{1}{6}\sum_{i=1}^{6}$$

$$= \frac{21}{6} = \frac{7}{2} = 3.5$$

2) fair coin flipping: $X = \begin{cases} 0 & \text{if outcome is } T \\ 1 & \text{'' '' '' } H \end{cases}$

$$\Pr[X=0] = 1/2 = \Pr[X=1]$$

$$\mathbb{E}[X] = \sum_{i=0}^{1} i \Pr[X=i] = 1/2$$

3) general coin flipping: $X = \begin{cases} 0 & \text{if outcome is } T \\ 1 & \text{'' '' '' } H \end{cases}$

$$\Pr[X=1] = p,\ \Pr[X=0] = 1-p$$

$$\Rightarrow \mathbb{E}[X] = 0 \cdot (1-p) + 1 \cdot p = p$$

Theorem

Let $g(X)$ be a function of a discrete random variable X. Then $\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$.

Example die rolling

X = outcome of a die, squared

Y = outcome of a die

$X = Y^2 = g(Y)$

$Y = 1, X = 1$ $\quad$ $Y = 2, X = 4 = g(Y)$

$$\mathbb{E}[X] = 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 9 \cdot \frac{1}{6} + 16 \cdot \frac{1}{6} + 25 \cdot \frac{1}{6}$$

$$+ 36 \cdot \frac{1}{36} = \frac{1}{6} \cdot 91 \approx 15.16\ldots$$

Theorem

Let $g(X)$ be a function of a discrete random variable X. Then $\mathbb{E}[g(X)] = \sum_x g(x)p_X(x)$.

For a random variable X we define:

- **Mean**: $m_X \doteq \mathbb{E}[X]$
- **Variance**: $\sigma_X^2 \doteq \mathbb{E}[\overbrace{(X - m_X)}^{\text{random variable}}{}^2] = \mathbb{E}[X^2] - m_X^2 = \textbf{Var}[X]$

$$\mathbb{E}[X - \mathbb{E}[X]] =$$

$$= \mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X]] =$$

$$= \mathbb{E}[X] - \mathbb{E}[X] = 0$$

Theorem

Let $g(X)$ be a function of a discrete random variable X. Then $\mathbb{E}[g(X)] = \sum_x g(x)p_X(x)$.

For a random variable X we define:

- **Mean**: $m_X \doteq \mathbb{E}[X]$
- **Variance**: $\sigma_X^2 \doteq \mathbb{E}[(X - m_X)^2] = \mathbb{E}[X^2] - m_X^2 = \mathbf{Var}[X]$
- k-th moment: $\mathbb{E}[X^k]$

<u>Example</u> die rolling

Y = outcome of a die

mean : $m_Y = \mathbb{E}[Y] = 3.5$

variance $\sigma_Y^2 = \mathbb{E}[Y^2] - (m_Y)^2$

$$= \frac{91}{6} - \left(\frac{21}{6}\right)^2$$

$$= \frac{35}{12} \approx 2.916\ldots$$

For a vector valued r.v. $\mathbf{X} \in \mathbb{R}^n$

Expectation:

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

$$\vec{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}^T$$

$$m_{X_i} = \mathbb{E}[X_i]$$

For a vector valued r.v. $\mathbf{X} \in \mathbb{R}^n$

Expectation:

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

Instead of the variance, we have the *covariance matrix*:

$$\Sigma = \mathbb{E}[(\mathbf{X} - m_{\mathbf{X}})(\mathbf{X} - m_{\mathbf{X}})^T] = \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1,X_2} & \cdots & \sigma_{X_1,X_n} \\ \sigma_{X_2,X_1} & \sigma^2_{X_2} & \vdots & \sigma_{X_2,X_n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{X_n,X_1} & \sigma_{X_n,X_2} & \cdots & \sigma^2_{X_n} \end{bmatrix}$$

where

$$\sigma_{X_i,X_j} = \mathbf{Cov}(X_i, X_j) \doteq \mathbb{E}[(X_i - m_{X_i})(X_j - m_{X_j})]$$

For a vector valued r.v. $\mathbf{X} \in \mathbb{R}^n$

Expectation:

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

Instead of the variance, we have the *covariance matrix*:

$$\Sigma = \mathbb{E}[(\mathbf{X} - m_{\mathbf{X}})(\mathbf{X} - m_{\mathbf{X}})^T] = \begin{bmatrix} \sigma^2_{X_1} & \sigma_{X_1,X_2} & \cdots & \sigma_{X_1,X_n} \\ \sigma_{X_2,X_1} & \sigma^2_{X_2} & \vdots & \sigma_{X_2,X_n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{X_n,X_1} & \sigma_{X_n,X_2} & \cdots & \sigma^2_{X_n} \end{bmatrix}$$

where

$$\sigma_{X_i,X_j} \quad = \quad \mathbf{Cov}(X_i, X_j) \doteq \mathbb{E}[(X_i - m_{X_i})(X_j - m_{X_j})]$$

σ_{X_i,X_j} is the covariance of X_i and X_j

Theorem

If X_1, X_2 are independent then $\sigma_{X_1,X_2} = 0$.

•) $\vec{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$, $X_1, X_2, \dots, X_n$ are mutually independent

$$\Sigma = \begin{bmatrix} \sigma_{X_1}^2 & & 0 \\ & \sigma_{X_2}^2 & \\ 0 & & \ddots \\ & & & \sigma_{X_n}^2 \end{bmatrix}$$

•) If $\sigma_{X_1,X_2} = 0$ then X_1, X_2 are independent? NO!

Theorem

If X_1, X_2 are independent then $\sigma_{X_1,X_2} = 0$.

The other direction is not true!

Counterexample: $X_1 = \begin{cases} 1 & \text{with prob } 1/2 \\ -1 & \text{with prob } 1/2 \end{cases}$ $\quad \mathbb{E}[X_1] = 0$

$$X_2 = \begin{cases} 0 & \text{(with prob. 1)} \quad \text{if } X_1 = -1 \\ -1 & \text{with prob. } \frac{1}{2} \quad \text{if } X_1 = 1 \\ 1 & \text{with prob. } \frac{1}{2} \quad \text{if } X_1 = 1 \end{cases} \implies X_1, X_2 \text{ are not independent}$$

$$\mathbb{E}[X_2] = -1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} = 0$$

But:

$$\sigma_{X_1,X_2} = \mathbb{E}[(X_1 - m_{X_1})(X_2 - m_{X_2})]$$
$$= \mathbb{E}[X_1 X_2 - \overbrace{m_{X_1}}^{0} X_2 - X_1 \overbrace{m_{X_2}}^{0} + \overbrace{m_{X_1} m_{X_2}}^{0}]$$
$$= \mathbb{E}[X_1 X_2] = 0 + 1 \cdot \frac{1}{4} - 1 \cdot \frac{1}{4} = 0$$

Theorem (Properties of Mean, Variance, etc.)

- $\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$ (linearity of expectation)
- constants $\mathbf{Var}[aX + b] = a^2\mathbf{Var}[X]$
- $\mathbf{Var}[X_1 + X_2] = \mathbf{Var}[X_1] + \mathbf{Var}[X_2] + 2\sigma_{X_1,X_2}$

Corollary

If $\sigma_{X_1,X_2} = 0$ *then* $\mathbf{Var}[X_1 + X_2] = \mathbf{Var}[X_1] + \mathbf{Var}[X_2]$

<u>Corollary</u>

If X_1, X_2 (r.v.'s) are independent

$\Rightarrow \mathrm{Var}[X_1 + X_2] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2]$

Conditional Probability

Definition (Conditional probability)

A, B are events: $\mathbb{P}[A|B] \doteq \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$. Well defined only if $\mathbb{P}[B] > 0$.

Example (Relative frequency, convergence, and conditional probability)

Consider an event A. $X_1, \ldots, X_n$ that are independent and identically distributed (i.i.d.) random variables that are indicator functions:

$$X_i = \begin{cases} 1 & \text{if } A \text{ happens} \\ 0 & \text{otherwise} \end{cases}$$

$$X_i(z) = \begin{cases} 1 & \text{if } z \in A \\ 0 & \text{if } z \notin A \end{cases}$$

Example (Relative frequency, convergence, and conditional probability)
Consider an event A. $X_1, \ldots, X_n$ that are independent and identically distributed (i.i.d.) random variables that are indicator functions:

$$X_i(z) = \begin{cases} 1 & \text{if } z \in A \\ 0 & \text{if } z \notin A \end{cases}$$

$$S_n = \sum_{i=1}^{n} X_i$$

Relative frequency: $f_n(A) \doteq \frac{S_n}{n}$
(of event A)

[proportion (ratio) of the "trials" in which event A happened.]

Example (Relative frequency, convergence, and conditional probability)
Consider an event A. $X_1, \ldots, X_n$ that are independent and identically distributed (i.i.d.) random variables that are indicator functions:

$$X_i(z) = \begin{cases} 1 & \text{if } z \in A \\ 0 & \text{if } z \notin A \end{cases}$$

$$S_n = \sum_{i=1}^{n} X_i$$

Relative frequency: $f_n(A) \doteq \frac{S_n}{n}$

Example
Coin flips, event $A =$ "the result of the coin flip is head"

Each X_i is a **Bernoulli r.v.** of parameter p: $X_i \sim B(p)$

Bernoulli of parameter p

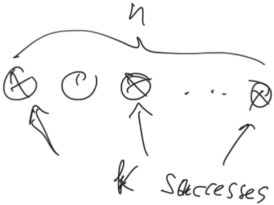
Each X_i is a **Bernoulli r.v.** of parameter p: $X_i \sim B(p)$

$$p = \mathbb{P}[X_i = 1] = \mathbb{P}[z \in A]$$

Then $S_n = \sum_{i=1}^{n} X_i$ is a **Binomial r.v.** of parameters n, p:

$$S_n \sim Bin(n, p)$$

$$\Pr[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}$$



Each X_i is a **Bernoulli r.v.** of parameter p: $X_i \sim B(p)$

$$p = \mathbb{P}[X_i = 1] = \mathbb{P}[z \in A]$$

Then $S_n = \sum_{i=1}^{n} X_i$ is a **Binomial r.v.** of parameters n, p:

$$S_n \sim Bin(n, p)$$

$$\mathbb{P}[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

$$\mathbb{E}[S_n] = np$$

$$Var[S_n] = np(1-p)$$

Each X_i is a **Bernoulli r.v.** of parameter p: $X_i \sim B(p)$

$$p = \mathbb{P}[X_i = 1] = \mathbb{P}[z \in A]$$

Then $S_n = \sum_{i=1}^{n} X_i$ is a **Binomial r.v.** of parameters n, p:

$$S_n \sim Bin(n, p)$$

$$\mathbb{P}[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

Then

$$\mathbb{E}[S_n] = np$$

Each X_i is a **Bernoulli r.v.** of parameter p: $X_i \sim B(p)$

$$p = \mathbb{P}[X_i = 1] = \mathbb{P}[z \in A]$$

Then $S_n = \sum_{i=1}^{n} X_i$ is a **Binomial r.v.** of parameters n, p:

$$S_n \sim Bin(n, p)$$

$$\mathbb{P}[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

Then

$$\mathbb{E}[S_n] = np$$

$$\mathbf{Var}[S_n] = np(1-p)$$

Exercise

Derive $\mathbb{E}[S_n]$ and $\mathbf{Var}[S_n]$.

Let's go back to the relative frequency $f_n(A) \doteq \frac{S_n}{n}$:

$$S_n \sim Bin(n, p)$$

$$\mathbb{E}[f_n(A)] = \mathbb{E}\left[\frac{S_n}{n}\right]$$

$$= \frac{1}{n}\mathbb{E}[S_n] = \frac{1}{n} \cdot np = p$$

Exercise

Derive $\mathbb{E}[S_n]$ and $\mathbf{Var}[S_n]$.

Let's go back to the relative frequency $f_n(A) \doteq \frac{S_n}{n}$:

$$\mathbb{E}[f_n(A)] = p$$

and

$$\mathrm{Var}[f_n(A)] = \mathrm{Var}\left[\frac{S_n}{n}\right]$$

$$= \left(\frac{1}{n}\right)^2 \mathrm{Var}[S_n] =$$

$$= \frac{1}{n^2} \, n p (1-p)$$

$$= \frac{p(1-p)}{n}$$

Exercise

Derive $\mathbb{E}[S_n]$ and $\mathbf{Var}[S_n]$.

Let's go back to the relative frequency $f_n(A) \doteq \frac{S_n}{n}$:

$$\mathbb{E}[f_n(A)] = p$$

and

$$\mathbf{Var}[f_n(A)] = \frac{p(1-p)}{n}$$

Theorem (Chebyshev's inequality)

Let X be a r.v. with $\mathbb{E}[X] = \mu$ and $\mathbf{Var}[X] = \sigma^2$. Then:

$$\mathbb{P}[|X - \mu| > \varepsilon] \leq \frac{\sigma^2}{\varepsilon^2}.$$

Therefore

r.v. (X) $\mu = \mathbb{E}[X]$

$$\mathbb{P}[|f_n(A) - p| > \varepsilon] \leq \frac{p(1-p)}{n\varepsilon^2}$$

Therefore

$$\mathbb{P}[|f_n(A) - p| > \varepsilon] \leq \frac{p(1-p)}{n\varepsilon^2}$$

and

$$\lim_{n \to +\infty} f_n(A) = p$$

Note: there are tighter bounds then Chebyshev's, like Chernoff's and Hoeffding's - we will see them later.

Intermission

Theorem (Law of Large Numbers)

Let X_i, $i = 1, \dots, n$ *be i.i.d. with* $\mathbb{E}[X_i] = \mu$ *and* $\mathbf{Var}[X] = \sigma^2 < +\infty$.

$\mathbb{E}[\bar{x}] = \mu$

Then

$$\lim_{n \to +\infty} \mathbb{P}\left[\left|\frac{1}{n}\sum X_i - \mu\right| > \varepsilon\right] = 0.$$

Intermission

Theorem (Law of Large Numbers)

Let X_i, $i = 1, \ldots, n$ *be i.i.d. with* $\mathbb{E}[X_i] = \mu$ *and* $\mathbf{Var}[X] = \sigma^2 < +\infty$.
Then

$$\lim_{n \to +\infty} \mathbb{P}\left[\left|\frac{1}{n}\sum X_i - \mu\right| > \varepsilon\right] = 0.$$

Note: See Jupyter notebook for an example.

Example (continue)

Remark 1:

$$\lim_{n \to +\infty} f_n(A) = \mathbb{P}[A]$$

Remark 2:

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]} = \lim_{n \to +\infty} \frac{f_n(A \cap B)}{f_n(B)}$$

$$\frac{f_n(A \cap B)}{f_n(B)} = \frac{S_n(A \cap B)}{S_n(B)}$$

it's the fraction of times $A \cap B$ happens among those in which B happens.

Computing Conditional Probabilities

Definition (Conditional probability)

A, B are events: $\mathbb{P}[A|B] \doteq \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$. Well defined only if $\mathbb{P}[B] > 0$.

Theorem (Bayes Rule)

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[B|A]\mathbb{P}[A]}{\mathbb{P}[B]}$$

Theorem (Law of Total Probability)

Let $C_1, C_2, \ldots, C_n$ *be a partition of* Ω*:*

- $\cup_{i=1}^n C_i = \Omega$
- $C_i \cap C_j = \emptyset$

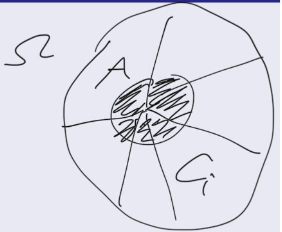
For all $A \subset \Omega$*:*

$$\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A|C_i]\mathbb{P}[C_i]$$

Theorem (Law of Total Probability)

Let $C_1, C_2, \ldots, C_n$ *be a partition of* Ω*:*

- $\cup_{i=1}^{n} C_i = \Omega$
- $C_i \cap C_j = \emptyset$



For all $A \subset \Omega$*:*

$$\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A|C_i]\mathbb{P}[C_i]$$

Example: $\mathbb{P}[B] = \mathbb{P}[B|A]\mathbb{P}[A] + \mathbb{P}[B|A^c]\mathbb{P}[A^c]$

opposite event ("not A")
complementary

Example

M = "have a rare disease", with $\mathbb{P}[M] = 10^{-9}$
T = "test for the disease is positive" with:

- $\mathbb{P}[T|M] = 0.99$ (1% false negatives)
- $\mathbb{P}[T|M^c] = 0.001$ (0.1% false positives)

If you test positive, what is the probability that you have the disease?

High : 29
Low : 13

Lunch: >> max {H, L}

Example

M = "have a rare disease", with $\mathbb{P}[M] = 10^{-9}$
T = "test for the disease is positive" with:

- $\mathbb{P}[T|M] = 0.99$ (1% false negatives)
- $\mathbb{P}[T|M^c] = 0.001$ (0.1% false positives)

If you test positive, what is the probability that you have the disease?

$= \Pr[T|M] \Pr[M] + \Pr[T|M^c] \Pr[M^c]$

$$\mathbb{P}[M|T] = \frac{\mathbb{P}[T|M]}{\mathbb{P}[T]}\mathbb{P}[M] = \frac{0.99 * 10^{-9}}{0.99 * 10^{-9} + 0.001(1 - 10^{-9})}$$

$$\approx \frac{1}{1 + 10^6} \approx 10^{-6}$$