Fundamentals of Machine Learning

BASICS OF PROBABILITY

Amit K Roy-Chowdhury

Acknowledgments: Partially adapted from Probability and Stochastic Processes: Yates and Goodman

What is Probability

- Describe phenomena that cannot be described with certainty because of the complexity of the underlying physical process.
- Different from your study of the deterministic sciences, e.g., the laws of classical mechanics.
- A number between 0 and 1.
- Probabilities are assigned based on observations, sometimes experience.
- Mathematical basis is in the theory of sets.

The terminology of set theory and probability

Set Algebra	Probability
Set	Event
Universal set	Sample space
Element	Outcome

Axioms of Probability

A probability measure $P[\cdot]$ is a function that maps events in the sample space to real numbers such that

Axiom 1 *For any event A, $P[A] \geq 0$.*

Axiom 2 $P[S] = 1$.

Axiom 3 *For any countable collection $A_1, A_2, \ldots$ of mutually exclusive events*

$$P\left[A_1 \cup A_2 \cup \cdots\right] = P\left[A_1\right] + P\left[A_2\right] + \cdots .$$

Basic Results in Probability

The probability measure $P[\cdot]$ satisfies

(a) $P[\phi] = 0$.

(b) $P[A^c] = 1 - P[A]$.

(c) For any A and B (not necessarily disjoint),

$$P[A \cup B] = P[A] + P[B] - P[A \cap B].$$

(d) If $A \subset B$, then $P[A] \leq P[B]$.

Conditional Probability

The conditional probability of the event A given the occurrence of the event B is

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

Total Probability

For an event space $\{B_1, B_2, \ldots, B_m\}$ with $P[B_i] > 0$ for all i,

$$P[A] = \sum_{i=1}^{m} P[A|B_i] P[B_i].$$

BAYES' THEOREM – VERY IMPORTANT

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

This is a very fundamental result that will arise throughout the course.

Practice Problem

A company has three machines B_1, B_2, and B_3 for making 1 kΩ resistors. It has been observed that 80% of resistors produced by B_1 are within 50 Ω of the nominal value. Machine B_2 produces 90% of resistors within 50 Ω of the nominal value. The percentage for machine B_3 is 60%. Each hour, machine B_1 produces 3000 resistors, B_2 produces 4000 resistors, and B_3 produces 3000 resistors. All of the resistors are mixed together at random in one bin and packed for shipment. What is the probability that the company ships a resistor that is within 50 Ω of the nominal value?

Solution

Let $A = \{$resistor is within 50 Ω of the nominal value$\}$. Using the resistor accuracy information to formulate a probability model, we write

$$P[A|B_1] = 0.8, \quad P[A|B_2] = 0.9, \quad P\left[A|B_3\right] = 0.6 \tag{1.29}$$

The production figures state that $3000 + 4000 + 3000 = 10{,}000$ resistors per hour are produced. The fraction from machine B_1 is $P[B_1] = 3000/10{,}000 = 0.3$. Similarly, $P[B_2] = 0.4$ and $P[B_3] = 0.3$. Now it is a simple matter to apply the law of total probability to find the accuracy probability for all resistors shipped by the company:

$$\begin{aligned} P[A] &= P[A|B_1]\,P[B_1] + P[A|B_2]\,P[B_2] + P\left[A|B_3\right]P\left[B_3\right] && (1.30)\\ &= (0.8)(0.3) + (0.9)(0.4) + (0.6)(0.3) = 0.78. && (1.31) \end{aligned}$$

For the whole factory, 78% of resistors are within 50 Ω of the nominal value.

Practice Problem – contd.

Find the probability that an acceptable resistor comes from machine B3.

Solution

Now we are given the event A that a resistor is within 50 Ω of the nominal value, and we need to find $P[B_3|A]$. Using Bayes' theorem, we have

$$P\left[B_3|A\right] = \frac{P\left[A|B_3\right] P\left[B_3\right]}{P\left[A\right]}.$$

Since all of the quantities we need are given in the problem description, our answer is P[B3] = 0.3, P[A]=0.78, P[A|B3] = 0.6

$$P\left[B_3|A\right] = (0.6)(0.3)/(0.78) = 0.23.$$

Similarly we obtain $P[B_1|A] = 0.31$ and $P[B_2|A] = 0.46$. Of all resistors within 50 Ω of the nominal value, only 23% come from machine B_3 (even though this machine produces 30% of all resistors). Machine B_1 produces 31% of the resistors that meet the 50 Ω criterion and machine B_2 produces 46% of them.

Independent Events

Events A *and* B *are* independent *if and only if*

$$P[AB] = P[A]\,P[B].$$

Independent vs Disjoint

Keep in mind that **independent and disjoint are *not* synonyms**.

In some contexts these words can have similar meanings, but this is not the case in probability. Disjoint events have no outcomes in common and therefore $P[AB] = 0$. In most situations independent events are not disjoint! Exceptions occur only when $P[A] = 0$ or $P[B] = 0$. When we have to calculate probabilities, knowledge that events A and B are *disjoint* is very helpful.

Practice Problem

A short-circuit tester has a red light to indicate that there is a short circuit and a green light to indicate that there is no short circuit. Consider an experiment consisting of a sequence of three tests. In each test the observation is the color of the light that is on at the end of a test. An outcome of the experiment is a sequence of red (r) and green (g) lights. We can denote each outcome by a three-letter word such as rgr, the outcome that the first and third lights were red but the second light was green. We denote the event that light n was red or green by R_n or G_n. The event $R_2 = \{grg, grr, rrg, rrr\}$. We can also denote an outcome as an intersection of events R_i and G_j. For example, the event $R_1G_2R_3$ is the set containing the single outcome $\{rgr\}$.

Practice Problem

Suppose that for the three lights (in previous slide) each outcome (a sequence of three lights, each either red or green) is equally likely. Are the events R_2 that the second light was red and G_2 that the second light was green independent? Are the events R_1 and R_2 independent?

Solution

Each element of the sample space

$$S = \{rrr, rrg, rgr, rgg, grr, grg, ggr, ggg\}$$

has probability $1/8$. Each of the events

$$R_2 = \{rrr, rrg, grr, grg\} \quad \text{and} \quad G_2 = \{rgr, rgg, ggr, ggg\}$$

contains four outcomes so $P[R_2] = P[G_2] = 4/8$. However, $R_2 \cap G_2 = \phi$ and $P[R_2 G_2] = 0$. That is, R_2 and G_2 must be disjoint because the second light cannot be both red and green. Since $P[R_2 G_2] \neq P[R_2]P[G_2]$, R_2 and G_2 are not independent. Learning whether or not the event G_2 (second light green) occurs drastically affects our knowledge of whether or not the event R_2 occurs. Each of the events $R_1 = \{rgg, rgr, rrg, rrr\}$ and $R_2 = \{rrg, rrr, grg, grr\}$ has four outcomes so $P[R_1] = P[R_2] = 4/8$. In this case, the intersection $R_1 \cap R_2 = \{rrg, rrr\}$ has probability $P[R_1 R_2] = 2/8$. Since $P[R_1 R_2] = P[R_1]P[R_2]$, events R_1 and R_2 are independent. Learning whether or not the event R_2 (second light red) occurs does not affect our knowledge of whether or not the event R_1 (first light red) occurs.

Three Independent Events

A_1, A_2, *and* A_3 *are* independent *if and only if*

(a) A_1 *and* A_2 *are independent,*

(b) A_2 *and* A_3 *are independent,*

(c) A_1 *and* A_3 *are independent,*

(d) $P[A_1 \cap A_2 \cap A_3] = P[A_1]P[A_2]P[A_3]$.

Random Variable

A random variable *consists of an experiment with a probability measure* $P[\cdot]$ *defined on a sample space* S *and a function that assigns a real number to each outcome in the sample space of the experiment.*

Discrete Random Variable

X *is a* discrete *random variable if the range of* X *is a countable set*

$$S_X = \{x_1, x_2, \ldots\}.$$

Probability Mass Function

The probability mass function *(PMF) of the discrete random variable* X *is*

$$P_X(x) = P[X = x]$$

Bernoulli Random Variable

X *is a* Bernoulli (p) *random variable if the PMF of* X *has the form*

$$P_X(x) = \begin{cases} 1-p & x=0 \\ p & x=1 \\ 0 & \text{otherwise} \end{cases}$$

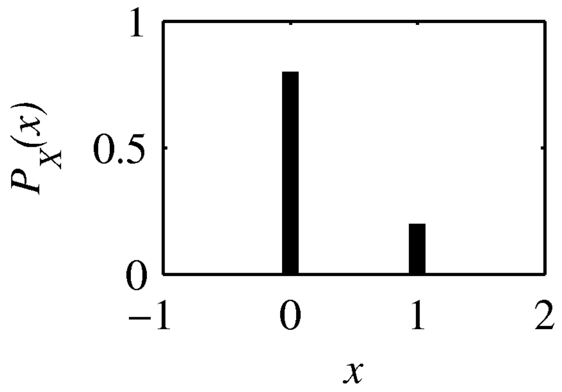
where the parameter p *is in the range* $0 < p < 1$.

Bernoulli Random Variable - Example

Suppose that a sample is rejected with probability p. Let X be the number of rejected samples in one test. X is a Bernoulli random variable.

Bernoulli Random Variable - Example

If there is a 0.2 probability of a reject,



$$P_X(x) = \begin{cases} 0.8 & x = 0 \\ 0.2 & x = 1 \\ 0 & \text{otherwise} \end{cases}$$

k rejects in n trials?

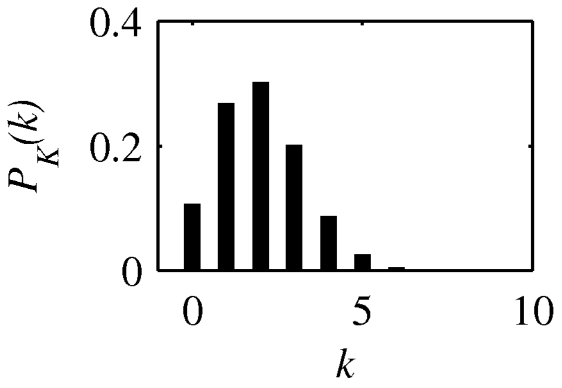
Suppose we test n circuits and each circuit is rejected with probability p independent of the results of other tests. Let K equal the number of rejects in the n tests. Find the PMF $P_K(k)$.

$$P_K(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Binomial Random Variable

Binomial Random Variable

If there is a 0.2 probability of a reject and we perform 10 tests,



$$P_K(k) = \binom{10}{k}(0.2)^k(0.8)^{10-k}.$$

Multinomial Distribution

A bag contains 8 red balls, 3 yellow balls, and 9 white balls. N = 6 balls are randomly selected **with replacement**. What is the probability that 2 are red, 1 is yellow and 3 are white?

W_i is the random variable denoting the number of balls of color i.

P(W_1=2, W_2=1, W_3=3) =

$$P(W_1 = n_1, ..., W_1 = n_k \mid N, \theta_1, .., \theta_k) = \frac{N!}{n_1!n_2!..n_k!} \theta_1^{n_1} \theta_2^{n_2} .. \theta_k^{n_k}$$

$$\sum_{i=1}^{k} n_i = N \qquad \sum_{i=1}^{k} \theta_i = 1$$

What happens if selection is **without** replacement?

Categorical Distribution

distribution over a finite set of labels, $y \in \{1, \ldots, C\}$

$$p(y = c|\boldsymbol{\theta}) = \theta_c$$

$$\mathrm{Cat}(y|\boldsymbol{\theta}) \triangleq \prod_{c=1}^{C} \theta_c^{\mathbb{I}(y=c)}$$

$$0 \leq \theta_c \leq 1 \qquad \sum_{c=1}^{C} \theta_c = 1$$

Roll a C-sided dice N times. y is the vector that counts the number of times each face shows up.

$$y_c = N_c \triangleq \sum_{n=1}^{N} \mathbb{I}(y_n = c)$$

Distribution of y is multinomial $\mathcal{M}(\boldsymbol{y}|N, \boldsymbol{\theta}) \triangleq \binom{N}{y_1 \ldots y_C} \prod_{c=1}^{C} \theta_c^{y_c} = \binom{N}{N_1 \ldots N_C} \prod_{c=1}^{C} \theta_c^{N_c}$ What happens when N=1?

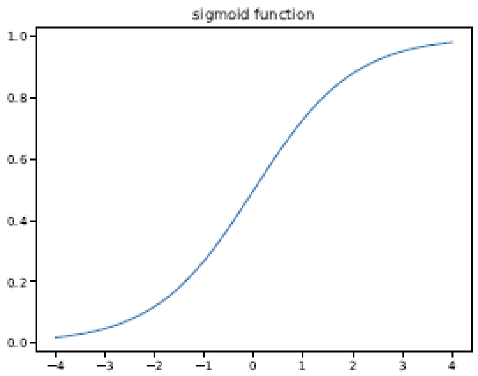
Why is the categorical distribution important? Think about the output of an ML model: $\mathrm{Cat}(y|f(\boldsymbol{x}; \boldsymbol{\theta}))$

Sigmoid (Logistic) Function

Predict binary random variable y given inputs x.

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y|f(\boldsymbol{x}; \boldsymbol{\theta}))$$

$$p(y|\boldsymbol{x}, \boldsymbol{\theta}) = \mathrm{Ber}(y|\sigma(f(\boldsymbol{x}; \boldsymbol{\theta})))$$



Logistic/Sigmoid function

$$\sigma(a) \triangleq \frac{1}{1+e^{-a}} = \frac{e^a}{1+e^a}$$

$$p(y=1|\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{1+e^{-a}} = \frac{e^a}{1+e^a} = \sigma(a)$$

$$p(y=0|\boldsymbol{x}, \boldsymbol{\theta}) = 1 - \frac{1}{1+e^{-a}} = \frac{e^{-a}}{1+e^{-a}} = \frac{1}{1+e^a} = \sigma(-a)$$

Logit function

$$a = \mathrm{logit}(p) = \sigma^{-1}(p) \triangleq \log\left(\frac{p}{1-p}\right)$$

log odds

$$\log\left(\frac{p}{1-p}\right) = \log\left(\frac{e^a}{1+e^a}\frac{1+e^a}{1}\right) = \log(e^a) = a$$

Will study logistic regression later

Cumulative Distribution Function

The cumulative distribution function *(CDF) of random variable* X *is*

$$F_X(x) = P[X \leq x].$$

Cumulative Distribution Function - Properties

For any random variable X,

(a) $F_X(-\infty) = 0$

(b) $F_X(\infty) = 1$

(c) $P[x_1 < X \le x_2] = F_X(x_2) - F_X(x_1)$

Continuous Random Variable

X *is a* continuous random variable *if the CDF* $F_X(x)$ *is a continuous function.*

Probability Density Function

The probability density function *(PDF) of a continuous random variable* X *is*

$$f_X(x) = \frac{dF_X(x)}{dx}.$$

Probability Density Function - Properties

For a continuous random variable X with PDF $f_X(x)$,

(a) $f_X(x) \geq 0$ for all x,

(b) $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du$,

(c) $\int_{-\infty}^{\infty} f_X(x)\, dx = 1$.

$$\Rightarrow \quad P\left[x_1 < X \leq x_2\right] = \int_{x_1}^{x_2} f_X(x)\, dx.$$

Expectation

The expected value *of* X *is*

$$E[X] = \mu_X = \sum_{x \in S_X} x P_X(x).$$

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx.$$

Variance

The variance *of random variable* X *is*

$$\mathrm{Var}[X] = E\left[(X - \mu_X)^2\right].$$

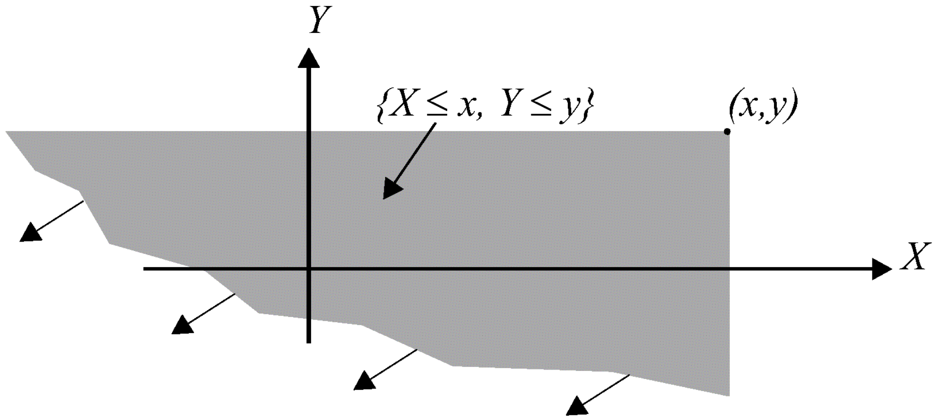
$$\mathrm{Var}[X] = E\left[X^2\right] - \mu_X^2 = E\left[X^2\right] - (E[X])^2$$

$$\mathrm{Var}[aX + b] = a^2\,\mathrm{Var}[X]$$

Multiple Random Variables

- Joint CDF, PDF
- Marginals

Joint CDF



The area of the (X, Y) plane corresponding to the joint cumulative distribution function $F_{X,Y}(x, y)$.

$$F_{X,Y}(x, y) = P[X \leq x, Y \leq y].$$

Joint PDF

The joint PDF of the continuous random variables X *and* Y *is a function* $f_{X,Y}(x, y)$ *with the property*

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v)\, dv\, du.$$

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x\, \partial y}$$

Marginal PDF

If X and Y are random variables with joint PDF $f_{X,Y}(x, y)$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dx.$$

Expectations, Covariance and Correlation

Expectation (function of random variable)

For random variables X and Y, the expected value of $W = g(X, Y)$ is

Discrete: $$E[W] = \sum_{x \in S_X} \sum_{y \in S_Y} g(x, y) P_{X,Y}(x, y)$$

Continuous: $$E[W] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\, dx\, dy.$$

Covariance and Correlation

The covariance *of two random variables* X *and* Y *is*

$$\mathrm{Cov}\left[X, Y\right] = E\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right].$$

The correlation *of* X *and* Y *is* $r_{X,Y} = E[XY]$

Uncorrelated and Orthogonal RVs

Random variables X *and* Y *are* orthogonal *if* $r_{X,Y} = 0$.

Random variables X *and* Y *are* uncorrelated *if* $\mathrm{Cov}[X, Y] = 0$.

When are they the same?

Correlation Coefficient

The correlation coefficient *of two random variables* X *and* Y *is*

$$\rho_{X,Y} = \frac{\operatorname{Cov}[X,Y]}{\sqrt{\operatorname{Var}[X]\operatorname{Var}[Y]}} = \frac{\operatorname{Cov}[X,Y]}{\sigma_X \sigma_Y}.$$

$$-1 \leq \rho_{X,Y} \leq 1.$$

Conditional Expected Value

The conditional expected value $E[X|Y]$ *is a function of random variable* Y *such that if* $Y = y$ *then* $E[X|Y] = E[X|Y = y]$.

Random Vectors

Expectation of a Random Vector

The expected value of a random vector $\mathbf{X}$ *is a column vector*

$$E[\mathbf{X}] = \boldsymbol{\mu}_{\mathbf{X}} = \begin{bmatrix} E[X_1] & E[X_2] & \cdots & E[X_n] \end{bmatrix}'.$$

Correlation Matrix

The correlation of a random vector $\mathbf{X}$ *is an* $n \times n$ *matrix* $\mathbf{R_X}$ *with* i, j*th element* $R_X(i, j) = E[X_i X_j]$. *In vector noation,*

$$\mathbf{R_X} = E\left[\mathbf{X}\mathbf{X}'\right].$$

Covariance Matrix

The covariance of a random vector $\mathbf{X}$ *is an* $n \times n$ *matrix* $\mathbf{C_X}$ *with components* $C_X(i, j) = \mathrm{Cov}[X_i, X_j]$. *In vector notation,*

$$\mathbf{C_X} = E\left[(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})(\mathbf{X} - \boldsymbol{\mu}_\mathbf{X})'\right]$$

Relation between Correlation and Covariance Matrices

For a random vector $\mathbf{X}$ with correlation matrix $\mathbf{R_X}$, covariance matrix $\mathbf{C_X}$, and vector expected value $\boldsymbol{\mu}_\mathbf{X}$,

$$\mathbf{C_X} = \mathbf{R_X} - \boldsymbol{\mu}_\mathbf{X}\boldsymbol{\mu}'_\mathbf{X}.$$