15-388/688 - Practical Data Science: Basic probability

J. Zico Kolter
Carnegie Mellon University
Fall 2019

Outline

Probability in data science

Basic rules of probability

Some common distributions

Outline

Probability in data science

Basic rules of probability

Some common distributions

Basic probability and statistics

Thus far, in our discussion of machine learning, we have largely avoided any talk of probability

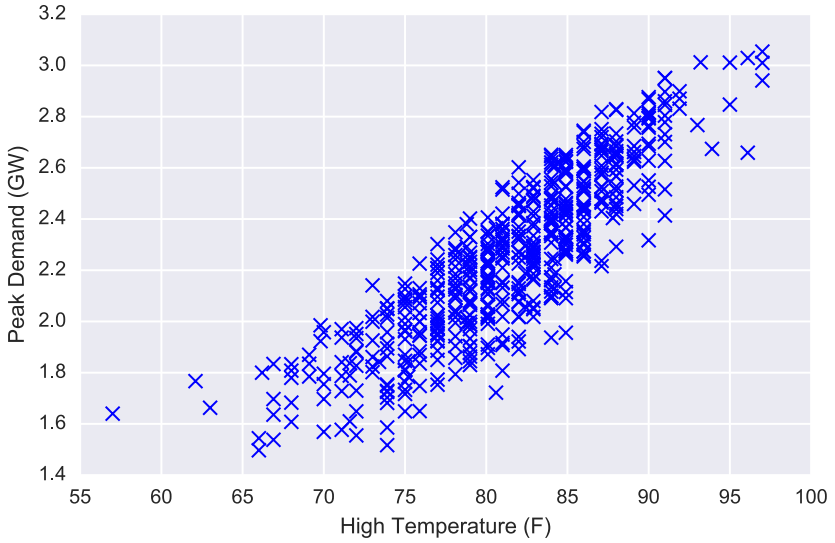
This won't be the case any longer, understanding and modeling probabilities is a crucial component of data science (and machine learning)

For the purposes of this course: statistics = probability + data

Probability and uncertainty in data science

In many prediction tasks, we never expect to be able to achieve perfect accuracy (there is some inherent randomness at the level we can observe the data)

In these situations, it is important to understand the uncertainty associated with our predictions



Outline

Probability in data science

Basic rules of probability

Some common distributions

Random variables

A random variable (informally) is a variable whose value is not initial known

Instead, these variables can take on different values (including a possibly infinite number), and must take on exactly one of these values, each with an associated probability, which all together sum to one

$$\text{“Weather” takes values } \{\text{sunny}, \text{rainy}, \text{cloudy}, \text{snowy}\}$$
$$p(\text{Weather} = \text{ sunny}) = 0.3$$
$$p(\text{Weather} = \text{rainy}) = 0.2$$
$$\ldots$$

Slightly different notation for continuous random variables, which we will discuss shortly

Notation for random variables

In this lecture, we use upper case letters, X to denote random variables

For a random variable X taking values $\{1,2,3\}$

$$p(X) = \begin{cases} 1: & 0.1 \\ 2: & 0.5 \\ 3: & 0.4 \end{cases}$$

represents a mapping from values to probabilities numbers that sum to one (odd notation, would be better to use p_X, but this is not common)

Conversely, we will use lower case x to denote a specific *value* of X (i.e., for above example $x \in \{1,2,3\}$), and $p(X = x)$ or just $p(x)$ refers to a *number* (the corresponding entry of $p(X)$)

Examples of probability notation

Given two random variables: X_1 with values in $\{1,2,3\}$ and X_2 with values in $\{1,2\}$:

- $p(X_1, X_2)$ refers to the joint distribution, i.e., a set of 6 possible values for each setting of variables, i.e. a dictionary mapping $(1,1), (1,2), (2,1), \ldots$ to corresponding probabilities)
- $p(x_1, x_2)$ is a number: probability that $X_1 = x_1$ and $X_2 = x_2$
- $p(X_1, x_2)$ is a set of 3 values, the probabilities for all values of X_1 for the given value $X_2 = x_2$, i.e., it is a dictionary mapping $0,1,2$ to numbers (note: not probability distribution, it will not sum to one)

We generally call all of these terms factors (dictionaries mapping values to numbers, even if they do not sum to one)

Example: weather and cavity

Let $\mathrm{Weather}$ denote a random variable taking on values in $\{\mathrm{sunny}, \mathrm{rainy}, \mathrm{cloudy}\}$ and Cavity a random variables taking on values in $\{\mathrm{yes}, \mathrm{no}\}$

$$P(\mathrm{Weather}, \mathrm{Cavity}) = \begin{cases} \mathrm{sunny}, \mathrm{yes} & 0.07 \\ \mathrm{sunny}, \mathrm{no} & 0.63 \\ \mathrm{rainy}, \mathrm{yes} & 0.02 \\ \mathrm{rainy}, \mathrm{no} & 0.18 \\ \mathrm{cloudy}, \mathrm{yes} & 0.01 \\ \mathrm{cloudy}, \mathrm{no} & 0.09 \end{cases}$$

$$p(\mathrm{sunny}, \mathrm{yes}) = 0.07$$

$$p(\mathrm{Weather}, \mathrm{yes}) = \begin{cases} \mathrm{sunny} & 0.07 \\ \mathrm{rainy} & 0.02 \\ \mathrm{cloudy} & 0.01 \end{cases}$$

Operations on probabilities/factors

We can perform operations on probabilities/factors by performing the operation on every corresponding value in the probabilities/factors

For example, given three random variables X_1, X_2, X_3:

$$p(X_1, X_2) \ \langle \text{op} \rangle \ p(X_2, X_3)$$

denotes a factor over X_1, X_2, X_3 (i.e., a dictionary over all possible combinations of values these three random variables can take), where the value for x_1, x_2, x_3 is given by

$$p(x_1, x_2) \ \langle \text{op} \rangle \ p(x_2, x_3)$$

Conditional probability

The **conditional probability** $p(X_1|X_2)$ (the conditional probability of X_1 given X_2) is defined as

$$p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)}$$

Can also be written $p(X_1, X_2) = p(X_1|X_2)p(X_2)$

Marginalization

For random variables X_1, X_2 with joint distribution $p(X_1, X_2)$

$$p(X_1) = \sum_{x_2} p(X_1, x_2) = \sum_{x_2} p(X_1|x_2)p(x_2)$$

Generalizes to joint distributions over multiple random variables

$$p(X_1, \dots, X_i) = \sum_{x_{i+1}, \dots, x_n} p(X_1, \dots, X_i, x_{i+1}, \dots, x_n)$$

For p to be a probability distribution, the marginalization over *all* variables must be one

$$\sum_{x_1, \dots, x_n} p(x_1, \dots, x_n) = 1$$

Bayes' rule

A straightforward manipulation of probabilities:

$$p(X_1|X_2) = \frac{p(X_1, X_2)}{p(X_2)} = \frac{p(X_2|X_1)p(X_1)}{p(X_2)} = \frac{p(X_2|X_1)p(X_1)}{\sum_{x_1} p(X_2|x_1)\, p(x_1)}$$

Poll: I want to know if I have come with with a rate strain of flu (occurring in only 1/10,000 people). There is an "accurate" test for the flu (if I have the flu, it will tell me I have 99% of the time, and if I do not have it, it will tell me I do not have it 99% of the time). I go to the doctor and test positive. What is the probability I have the this flu?

Bayes' rule

BBC Sign in | News | Sport | Weather | Shop | Earth | Travel

NEWS

Home | Video | World | US & Canada | UK | Business | Tech | Science | Stories

Magazine

Do doctors understand test results?

By William Kremer
BBC World Service

7 July 2014 Share

In one session, almost half the group of 160 gynaecologists responded that the woman's chance of having cancer was nine in 10. Only 21% said that the figure was one in 10 - which is the correct answer. That's a worse result than if the doctors had been answering at random.

The fact that 90% of women with breast cancer get a positive result from a mammogram doesn't mean that 90% of women with positive results have breast cancer. The high false alarm rate, combined with the disease's prevalence of 1%, means that roughly nine out of 10 women with a worrying mammogram don't actually have breast cancer.

Independence

We say that random variables X_1 and X_2 are **(marginally) independent** if their joint distribution is the product of their marginals

$$p(X_1, X_2) = p(X_1)p(X_2)$$

Equivalently, can also be stated as the condition that

$$p(X_1|X_2) \left(= \frac{p(X_1, X_2)}{p(X_2)} = \frac{p(X_1)p(X_2)}{p(X_2)}\right) = p(X_1)$$

$$(\text{and similarly}) \quad p(X_2|X_1) = p(X_2)$$

Poll: Weather and cavity

Are the weather and cavity random variables independent?

$$P(\text{Weather}, \text{Cavity}) = \begin{cases} \text{sunny}, \text{yes} & 0.07 \\ \text{sunny}, \text{no} & 0.63 \\ \text{rainy}, \text{yes} & 0.02 \\ \text{rainy}, \text{no} & 0.18 \\ \text{cloudy}, \text{yes} & 0.01 \\ \text{cloudy}, \text{no} & 0.09 \end{cases}$$

Conditional independence

We say that random variables X_1 and X_2 are **conditionally independent given** X_3, if

$$p(X_1, X_2|X_3) = p(X_1|X_3)p(X_2|X_3)$$

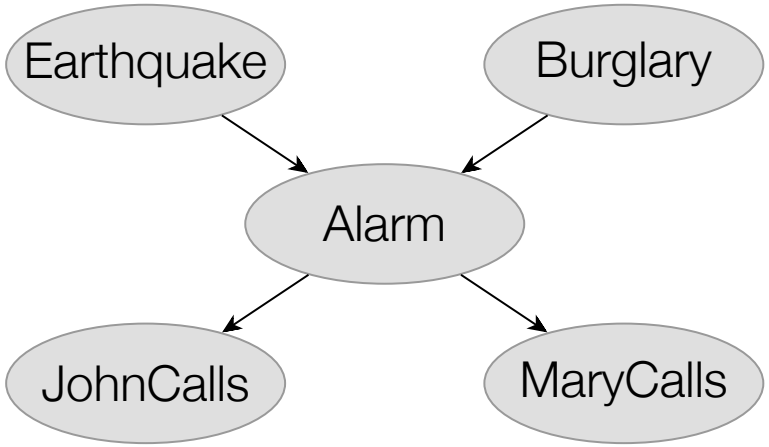
Again, can be equivalently written:

$$p(X_1|X_2, X_3) \left(= \frac{p(X_1, X_2|X_3)}{p(X_2|X_3)} = \frac{p(X_1|X_3)p(X_2|X_3)}{p(X_2|X_3)} \right) = p(X_1|X_3)$$

And similarly $p(X_2|X_1, X_3) = p(X_2|X_3)$

Marginal and conditional independence

Important: Marginal independence does not imply conditional independence or vice versa



$$P(\text{Earthquake}|\text{Burglary}) = P(\text{Earthquake}) \text{ but}$$
$$P(\text{Earthquake}|\text{Burglary}, \text{Alarm}) \neq P(\text{Earthquake}|\text{Alarm})$$

$$P(\text{JohnCalls}|\text{MaryCalls}, \text{Alarm}) = P(\text{JohnCalls}|\text{Alarm}) \text{ but}$$
$$P(\text{JohnCalls}|\text{MaryCalls}) \neq P(\text{JohnCalls})$$

Expectation

The expectation of a random variable is denoted:

$$\mathbf{E}[X] = \sum_x x \cdot p(x)$$

where we use upper case X to emphasize that this is a function of the entire random variable (but unlike $p(X)$ is a number)

Note that this only makes sense when the values that the random variable takes on are *numerical* (i.e., We can't ask for the expectation of the random variable "Weather")

Also generalizes to *conditional expectation:*

$$\mathbf{E}[X_1|x_2] = \sum_{x_1} x_1 \cdot p(x_1|x_2)$$

Rules of expectation

Expectation of sum is always equal to sum of expectations (even when variables are not independent):

$$\begin{aligned}
\mathbf{E}[X_1 + X_2] &= \sum_{x_1, x_2} (x_1 + x_2) p(x_1, x_2) \\
&= \sum_{x_1} x_1 \sum_{x_2} p(x_1, x_2) + \sum_{x_2} x_2 \sum_{x_1} p(x_1, x_2) \\
&= \sum_{x_1} x_1 p(x_1) + \sum_{x_2} x_2 p(x_2) \\
&= \mathbf{E}[X_1] + \mathbf{E}[X_2]
\end{aligned}$$

Rules of expectation

If x_1, x_2 independent, expectation of products is product of expectations

$$\begin{aligned}
\mathbf{E}[X_1 X_2] &= \sum_{x_1, x_2} x_1 x_2 \, p(x_1, x_2) \\
&= \sum_{x_1, x_2} x_1 x_2 \, p(x_1) p(x_2) \\
&= \sum_{x_1} x_1 p(x_1) \sum_{x_2} x_2 p(x_2) \\
&= \mathbf{E}[X_1] \mathbf{E}[X_2]
\end{aligned}$$

Variance

Variance of a random variable is the expectation of the variable minus its expectation, squared

$$\begin{aligned}\mathbf{Var}[X] &= \mathbf{E}[(X - \mathbf{E}[X])^2] \left(= \sum_x (x - \mathbf{E}[x])^2 p(x)\right) \\ &= \mathbf{E}[X^2 - 2X\mathbf{E}[X] + \mathbf{E}[X]^2] \\ &= \mathbf{E}[X^2] - \mathbf{E}[X]^2\end{aligned}$$

Generalizes to covariance between two random variables

$$\begin{aligned}\mathbf{Cov}[X_1, X_2] &= \mathbf{E}[(X_1 - \mathbf{E}[X_1])(X_2 - \mathbf{E}[X_2])] \\ &= \mathbf{E}[X_1 X_2] - \mathbf{E}[X_1]\mathbf{E}[X_2]\end{aligned}$$

Infinite random variables

All the math above works the same for discrete random variables that can take on an infinite number of values (for those with some math background, I'm talking about *countably infinite* values here*)*

The only difference is that $p(X)$ (obviously) cannot be specified by an explicit dictionary mapping variable values to probabilities, need to specify a *function* that produces probabilities

To be a probability, we still must have $\sum_x p(x) = 1$

Example:

$$P(X = k) = \left(\frac{1}{2}\right)^k, \qquad k = 1, \ldots, \infty$$

Continuous random variables

For random variables taking on *continuous* values (we'll only consider real-valued distributions), we need some slightly different mechanisms

As with infinite discrete variables, the distribution $p(X)$ needs to be specified as a function: here is referred to as a **probability density function** (PDF) and it must *integrate* to one $\int_{\mathbb{R}} p(x)dx = 1$

For any interval (a, b), we have that $p(a \leq x \leq b) = \int_a^b p(x)dx$ (with similar generalization to multi-dimensional random variables)

Can also be specified by their **cumulative distribution function** (CDF), $F(a) = p(x \leq a) = \int_\infty^a p(x)$

Outline

Probability in data science

Basic rules of probability

Some common distributions

Bernoulli distribution

A simple distribution over binary $\{0,1\}$ random variables

$$p(X=1;\phi)=\phi, \qquad P(X=0;\phi)=1-\phi$$

where $\phi \in [0,1]$ is the parameter that governs the distribution

Expectation is just $\mathbf{E}[x]=\phi$ (but not very common to refer to it this way, since this would imply that the $\{0,1\}$ terms are actual real-valued numbers)

Categorical distribution

This is the discrete distribution we've mainly considered so far, a distribute over finite discrete elements with each probability specified

Written generically as:

$$p(X = i; \phi) = \phi_i$$

where $\phi_1, \ldots \phi_k \in [0,1]$ are the parameters of the distribution (the probability of each random variable, must sum to one)

Note: we could actually parameterize just using $\phi_1, \ldots \phi_{k-1}$, since this would determine the last elements

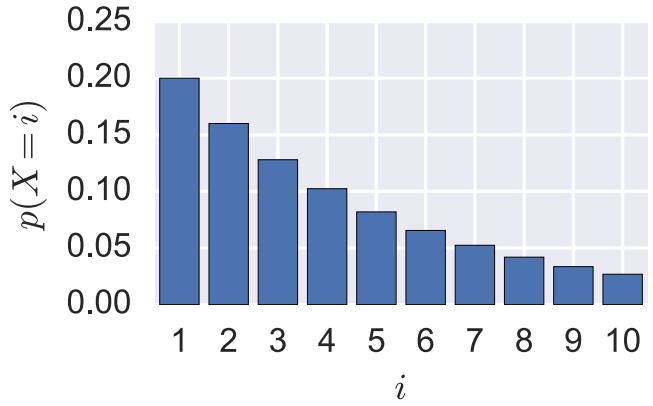
Unless the actual numerical value of the i's are relevant, it doesn't make sense to take expectations of a categorical random variable

Geometric distribution

The geometric distribution is an distribution over the positive integers, can be viewed as the number of Bernoulli trials needed before we get a "1"

$$p(X = i; \phi) = (1 - \phi)^{i-1}\phi, \qquad i = 1, \ldots, \infty$$

where $\phi \in [0,1]$ is parameter governing distribution (also $\mathbf{E}[X] = 1/\phi$)



$\phi = 0.2$

Note: easy to check that

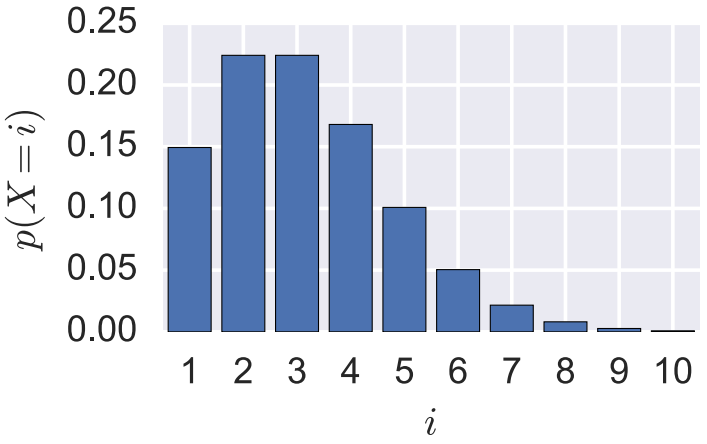
$$\sum_{i=1}^{\infty} p(X = i) = \phi \sum_{i=1}^{\infty} (1 - \phi)^{i-1} = \phi \cdot \frac{1}{1 - (1 - \phi)} = 1$$

Poisson distribution

Distribution over non-negative integers, popular for modeling number of times an event occurs within some interval

$$P(X = i; \lambda) = \frac{\lambda^i e^{-\lambda}}{i!}, \qquad i = 0, \ldots, \infty$$

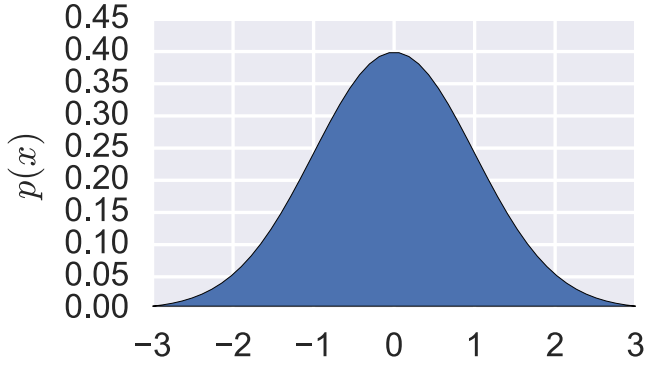
where $\lambda \in \mathbb{R}$ is parameter governing distribution (also $\mathbf{E}[X] = \lambda$)



$\lambda = 3$

Gaussian distribution

Distribution over real-valued numbers, empirically the most common distribution in all of data science (not in data itself, necessarily, but for people applying data science), the standard “bell curve”:



$$\mu = 0$$
$$\sigma^2 = 1$$

Probability density function:

$$p(x; \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \equiv \mathcal{N}(x; \mu, \sigma^2)$$

with parameters $\mu \in \mathbb{R}$ (mean) and $\sigma^2 \in \mathbb{R}_+$ (variance)

Multivariate Gaussians

The Gaussian distribution is one of the few distributions that generalizes nicely to higher dimensions

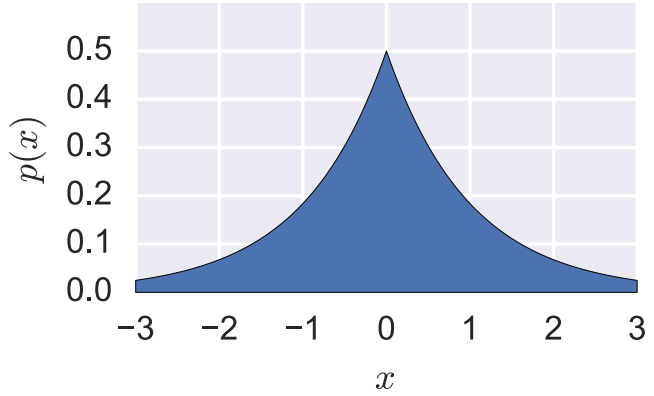
We'll discuss this in much more detail when we talk about anomaly detection and the mixture of Gaussians model, but for now, just know that we can also write a distribution over random *vectors* $x \in \mathbb{R}^n$

$$p(x; \mu, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\Big(-(x-\mu)^T \Sigma^{-1} (x-\mu)\Big)$$

where $\mu \in \mathbb{R}^n$ is mean and $\Sigma \in \mathbb{R}^{n \times n}$ is *covariance matrix,* and $|\cdot|$ denotes the determinant of a matrix

Laplace distribution

Like a Gaussian but with absolute instead of squared difference, gives the distribution (relatively) “heavy tails”



$\mu = 0$
$b = 1$

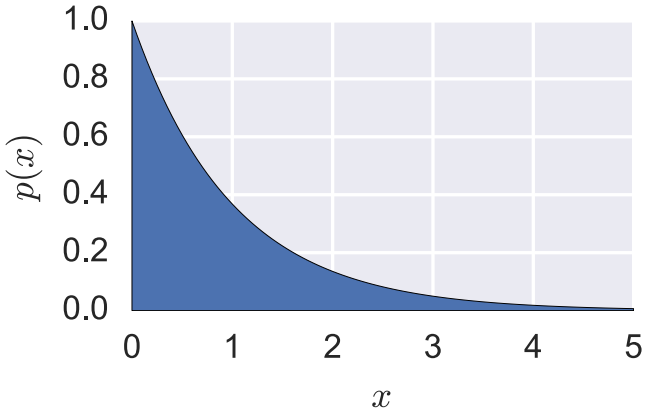
Probability density function:

$$p(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x - \mu|}{b}\right)$$

with parameters μ (mean), b (variance is $2b^2$)

Exponential distribution

A one-sided Laplace distribution, often used to model arrival times



$\lambda = 1$

Probability density function:

$$p(x; \lambda) = \lambda \exp(-\lambda x)$$

with parameter $\lambda \in \mathbb{R}_+$ (mean/variance $\mathbf{E}[X] = 1/\lambda, \mathbf{Var}[x] = 1/\lambda^2$)

Some additional examples

Student's t distribution – distribution governing estimation of normal distribution from finite samples, commonly used in hypothesis testing

χ^2 (chi-squared) distribution – distribution of Gaussian variable squared, also used in hypothesis testing

Cauchy distribution – very heavy tailed distribution, to the point that variables have undefined expectation (the associated integral is undefined)