CS246 (Winter 2015) Mining Massive Data Sets

Probability reminders

Sammy El Ghazzal, Jean-Yves Stephan

Disclaimer These notes may contain typos, mistakes or confusing points. Please contact jstephan@stanford.edu so that we can improve them for next year.

1 Definition: a few reminders

Definition (Sample space, Event space, Probability measure)**.**
This definition contains the basic definitions of probability theory:

- Sample space (usually denoted as Ω): the set of all possible outcomes.
- Event space (usually denoted as $\mathcal{F}$): a family of subsets of Ω (possibly all subsets of Ω).
- Probability Measure Function P: a function that goes from $\mathcal{F}$ to $\mathbb{R}$. It must have the following properties:
  1. $\mathbb{P}(\Omega) = 1$.
  2. $\forall A \in \mathcal{F},\ 0 \leq \mathbb{P}(A) \leq 1$.
  3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.
  4. For a set of disjoint events $A_1, \ldots, A_p$:

$$\mathbb{P}\left(\bigcup_{1 \leq i \leq p} A_i\right) = \sum_{i=1}^{p} \mathbb{P}(A_i).$$

Proposition (Union bound)**.**
Let A and B be two events. As we have seen, it holds that:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

and in particular (the following formula is referred to as the *Union Bound*):

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B),$$

and more generally if $E_1, \ldots, E_n$ are events:

$$\mathbb{P}\left(\bigcup_{1 \leq i \leq n} E_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(E_i).$$

Definition (Random variable)**.**
A random variable X is a function from the sample space to $\mathbb{R}$.

Definition (Cumulative distribution function)**.**
Let X be a random variable. Its cumulative distribution function (cdf) F is defined as:

$$F(x) = \mathbb{P}(X \leq x).$$

F is monotonically increasing and verifies:

$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to +\infty} F(x) = 1.$$

Definition (Probability density function)**.**
Let X be a continuous random variable X, the probability density function p_X of X is defined (when it exists) as the function such that:

$$dF(x) = p_X(x)dx,$$

where F is the cumulative distribution function of X.

p_X is non-negative and verifies:

$$\int_{\mathbb{R}} p_X(x)dx = 1.$$

Definition (Expectation)**.**
Let X be a continuous (resp. discrete) random variable. The expectation of X is defined as:

$$\mathbb{E}(X) = \int_{\mathbb{R}} p_X(x)x dx \quad \left(\text{resp. } \sum_x \mathbb{P}(X = x)\, x\right)$$

Proposition (Linearity of expectation)**.**
The expectation is linear, that is, if X and Y are random variables, then:

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) \text{ and } \forall a \in \mathbb{R},\ \mathbb{E}(aX) = a\mathbb{E}(X).$$

Definition (Variance)**.**
Let X be a random variable. The variance of X is defined as:

$$\text{var}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right) = \mathbb{E}\left(X^2\right) - \mathbb{E}(X)^2.$$

The standard deviation of X (also often denoted as σ_X) is then defined as:

$$\text{std}(X) = \sqrt{\text{var}(X)}.$$

Definition (Covariance)**.**
Let X and Y be two random variables. The covariance of X and Y is defined as:

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

In particular, note that:

$$\text{var}(X) = \text{cov}(X, X).$$

The correlation coefficient between X and Y is then defined as:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}.$$

The correlation can therefore be thought of as the "normalized" covariance.

Proposition (Variance of sum of random variables)**.**
Let X and Y be random variables. Then:

$$\text{var}(X+Y) = \text{var}(X) + 2\text{cov}(X,Y) + \text{var}(Y) \text{ and } \forall a \in \mathbb{R},\ \text{var}(aX) = a^2\text{var}(X).$$

Definition (Independence)**.**
Let X and Y be random variables. We say that X and Y are independent if and only if:

$$\forall U, V,\ \mathbb{P}\left(X \in U,\ Y \in V\right) = \mathbb{P}\left(X \in U\right)\mathbb{P}\left(Y \in V\right).$$

Proposition (Independence and covariance)**.**
Let X and Y be two random variables. If X and Y are independent, then:

$$\text{cov}(X,Y) = 0.$$

The opposite is not true in general.

Note that this result implies that if X and Y are independent random variables, then:

$$\text{var}(X+Y) = \text{var}(X) + \text{var}(Y).$$

Definition (Bayes rule)**.**
Let A and B be two events. Then:

$$\mathbb{P}\left(A \mid B\right) = \frac{\mathbb{P}\left(A \cap B\right)}{\mathbb{P}\left(B\right)}.$$

Proposition (Law of total probability)**.**
Let A be an event and B be a non-negative discrete random variable. Then:

$$\mathbb{P}\left(A\right) = \sum_{k=0}^{+\infty} \mathbb{P}\left(A \mid B = k\right)\mathbb{P}\left(B = k\right).$$

Definition (Indicator variable)**.**
Let A be an event. We define the indicator variable denoted as I_A or $\mathbb{1}_A$ as:

$$I_A = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

I_A has the following property:

$$\mathbb{E}\left(I_A\right) = \mathbb{P}\left(A\right).$$

2 Common distributions

Let us start with the common distributions of discrete random variables:

- $X \sim \mathcal{B}(p)$: X is a Bernoulli random variable with parameter p if and only if:

  $$\mathbb{P}\left(X = 1\right) = p \text{ and } \mathbb{P}\left(X = 0\right) = 1 - p.$$

  Then:

  $$\mathbb{E}\left(X\right) = p \text{ and } \text{var}(X) = p(1-p).$$

  Example Coin flip.

- $X \sim \mathcal{B}(n,p)$: X is a Binomial random variable with parameters n and p if and only if:

$$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

  Then:

$$\mathbb{E}(X) = np \text{ and } \mathrm{var}(X) = np(1-p).$$

  Example Number of heads obtained in n coin flips.

- $X \sim \mathcal{P}(\lambda)$: X is a Poisson random variable of parameter λ if and only if:

$$\mathbb{P}(X=k) = \frac{e^{-\lambda}\lambda^k}{k!}.$$

  Then:

$$\mathbb{E}(X) = \lambda \text{ and } \mathrm{var}(X) = \lambda.$$

  Example Number of people arriving in a queue.

The most common continous distribution that we will use is the Gaussian distribution: $X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if the probability density function of X is:

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Then:

$$\mathbb{E}(X) = \mu \text{ and } \mathrm{var}(X) = \sigma^2.$$

3 Common inequalities

Proposition (Markov inequality)**.**
Let X be a non-negative random variable. Then:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Proposition (Chebychev inequality)**.**
Let X be a random variable. Then:

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathrm{var}(X)}{a^2}.$$

Proposition (Hoeffding inequality)**.**
Let $X_1, \ldots, X_n$ be i.i.d. random variables such that: $\forall 1 \leq i \leq n,\ X_i \in [0,1]$. We denote by μ the expectation of the X_i's.

Then:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Proposition (Jensen inequality)**.**
Let ϕ be a convex function and X be a random variable. Then:

$$\mathbb{E}(\phi(X)) \geq \phi(\mathbb{E}(X)).$$

Proposition.
The following limit holds:

$$\forall x \in \mathbb{R},\ \lim_{n\to\infty} \left(1 - \frac{x}{n}\right)^n = e^{-x}.$$

Note that the following inequality also holds:

$$\forall x \in \mathbb{R},\ \left(1 - \frac{x}{n}\right)^n \leq e^{-x}.$$

Proposition (Stirling formula)**.**
An equivalent of $n!$ when n goes to infinity is:

$$n! \underset{n\to\infty}{\sim} \sqrt{2\pi n} e^{-n} n^n.$$

4 Maximum Likelihood Estimation

The method of maximum likelihood estimation (MLE) is a method that can be used to estimate the parameters of a model. The goal is to find the value of the parameter(s) that maximize the probability of the data sample (*i.e.* the likelihood) being observed (given the model).

For instance, if you assume that some dataset can be modeled as Gaussian and you want to estimate the parameters of the Gaussian (mean and variance), MLE is a good method to use and it will find the parameters that "fit" best the data.

Let us go through an example to explain the method: say we have n data points $(x_1, \ldots, x_n)$ drawn i.i.d. from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ and we want to estimate μ and σ^2 from these samples.

We compute the likelihood of the observations:

$$\begin{aligned}\mathcal{L}(\mu, \sigma) &= p(x_1, \ldots, x_n; \mu, \sigma) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},\end{aligned}$$

and the goal is now to maximize $\mathcal{L}$ with respect to the parameters μ and σ.

As we have a product and exponentials, it is in fact easier to maximize the log-likelood defined as:

$$\begin{aligned}\ell(\mu, \sigma) &= \log \mathcal{L}(\mu, \sigma) \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}.\end{aligned}$$

We now look for μ^* and σ^* by solving:

$$\begin{aligned}\frac{\partial \ell}{\partial \mu}(\mu^*, \sigma^*) &= 0 \\ \frac{\partial \ell}{\partial \sigma}(\mu^*, \sigma^*) &= 0.\end{aligned}$$

The first equation gives us:

$$\frac{\partial \ell}{\partial \mu}(\mu^*, \sigma^*) = -\sum_{i=1}^{n} \frac{x_i - \mu}{\sigma^2},$$

and therefore to make this derivative be 0, we need:

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The second equation gives:

$$\frac{\partial \ell}{\partial \sigma}(\mu^*, \sigma^*) = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{\sigma^3},$$

and therefore to make this derivative be 0, we need:

$$\sigma* = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu^*)^2},$$

which concludes the estimation of the parameters of the model.

5 Exercises

1. Let X be a random variable that takes non-negative integer values. Prove that:

$$\mathbb{E}(X) = \sum_{k=1}^{+\infty} \mathbb{P}(X \geq k).$$

Solution We compute:

$$\begin{aligned}
\sum_{k=1}^{+\infty} \mathbb{P}(X \geq k) &= \sum_{k=1}^{+\infty}\sum_{u=k}^{+\infty} \mathbb{P}(X = u) \\
&= \sum_{u=1}^{+\infty}\sum_{k=1}^{u} \mathbb{P}(X = u) \qquad \text{(Change order of summation)} \\
&= \sum_{u=1}^{+\infty} u\mathbb{P}(X = u) \\
&= \mathbb{E}(X).
\end{aligned}$$

2. (Birthday Paradox) Let us consider a room with $n \leq 365$ people. Compute the probability that two people in the room share the same birthday (for simplicity, we will consider that a year is 365 days).

Solution We start by computing the probability that no two people have the same birthday.

One way to solve the problem is to look at how many possibilities each person (taken in a fixed order) has: the first person will have 365 possible days, the second 364, ...

This gives:

$$\begin{aligned}
\mathbb{P}(\text{No two people have the same birthday}) &= \frac{365}{365} \times \frac{364}{365} \times \cdots \times \frac{365 - (n-1)}{365} \\
&= \frac{365!}{(365-n)! \times 365^n} \\
&= \frac{\binom{365}{n} \times n!}{365^n}
\end{aligned}$$

Therefore, the probability that two people share the same birthday is:

$$1 - \frac{\binom{365}{n} n!}{365^n}.$$

3. Show that if $X \sim \mathcal{P}(\lambda)$ and $Y \mid X = k \sim \mathcal{B}(k, p)$ then $Y \sim \mathcal{P}(\lambda p)$.

   Hint: Use the definitions of the Poisson and Binomial distributions, as well as the law of total probability.

   Solution We compute:

$$\begin{aligned}
\mathbb{P}(Y = k) &= \sum_{u=0}^{+\infty} \mathbb{P}(Y = k \mid X = u)\, \mathbb{P}(X = u) \\
&= \sum_{u=k}^{+\infty} \mathbb{P}(Y = k \mid X = u)\, \mathbb{P}(X = u) \\
&= \sum_{u=k}^{+\infty} \binom{u}{k} p^k (1-p)^{u-k} \frac{e^{-\lambda}\lambda^u}{u!} \\
&= \frac{p^k e^{-\lambda}}{k!} \sum_{u=k}^{+\infty} \frac{\lambda^u (1-p)^{u-k}}{(u-k)!} \\
&= \frac{p^k e^{-\lambda}}{k!} \sum_{u=k}^{+\infty} \frac{\lambda^{u-k}\lambda^k (1-p)^{u-k}}{(u-k)!} \\
&= \frac{(\lambda p)^k e^{-\lambda}}{k!} \sum_{u=k}^{+\infty} \frac{\lambda^{u-k} (1-p)^{u-k}}{(u-k)!} \\
&= \frac{(\lambda p)^k e^{-\lambda}}{k!} \sum_{q=0}^{+\infty} \frac{\lambda^q (1-p)^q}{q!} \\
&= \frac{(\lambda p)^k e^{-\lambda}}{k!} e^{\lambda(1-p)} \\
&= \frac{(\lambda p)^k e^{-\lambda p}}{k!},
\end{aligned}$$

   which proves the result.

4. Why do you need to assume that X be a non-negative random variable in Markov inequality (find an example with a random variable that is non-negative and the proposition does not hold) ?

   Solution Let X be a random variable that takes the values -1 and 1 with probability 0.5. Then $\mathbb{E}(X) = 0$ and if we take $a = 0.1$ then Markov inequality does not hold (0.5 on the left-hand side and 0 on the right hand-side).

5. Let us consider a setting with n birds and k boxes. Assume each bird picks one of the boxes uniformly at random and independently from the other birds.

   (a) Let X be the random variable representing the number of birds in box number 0. What are $\mathbb{E}(X)$ and $\mathrm{var}(X)$?

(b) Compute the probability p_0 that only birds i and $j > i$ (i and j are some fixed indices) end up in box 0 (*i.e.* these two birds are by themselves in box 0). From this result, compute the probability that birds i and j end up by themselves in the same box (not necessarily box 0).

(c) Assuming that $k = n - 2$, find a simple upper-bound on p_0.

Solution

(a) Let us denote by B_i the event that bird i goes to box 0. The B_i are i.i.d. random variables distributed as $\mathcal{B}(\frac{1}{k})$. We therefore compute by linearity of expectation:

$$\mathbb{E}(X) = \mathbb{E}\left(\sum_{i=1}^{n} B_i\right) = \sum_{i=1}^{n} \mathbb{E}(B_i) = \frac{n}{k}.$$

For the variance, you need to be a bit more careful as it is not a linear operator. But here, the choices are independent and therefore:

$$\mathrm{var}(X) = \mathrm{var}\left(\sum_{i=1}^{n} B_i\right) \underset{\text{Independence}}{=} \sum_{i=1}^{n} \mathrm{var}(B_i) = n \times \frac{1}{k}\left(1 - \frac{1}{k}\right).$$

(b) Let us call M_{ij} the event that birds i and $j > i$ end up in box 0 and no other bird goes to box 0. We have, by independence:

$$p_0 = \mathbb{P}(M_{ij}) = \frac{1}{k^2}\left(1 - \frac{1}{k}\right)^{n-2},$$

Let us call A_p the event that birds i and j end up in box p. The $(A_p)_{1 \le p \le k}$ are clearly disjoint and therefore:

$$\mathbb{P}(i \text{ and } j \text{ end up in the same box}) = \mathbb{P}\left(\bigcup_{1 \le p \le k} A_p\right) = \sum_{p=1}^{k} \underbrace{\mathbb{P}(A_p)}_{=p_0} = kp_0 = \frac{1}{k}\left(1 - \frac{1}{k}\right)^{n-2}.$$

(c) Using the common inequality:

$$\left(1 - \frac{1}{p}\right)^{p} \le e^{-1},$$

we compute:

$$p_0 \le \frac{1}{e(n-2)^2}.$$

6. Let us consider n points $(x_1, \ldots, x_n)$ drawn i.i.d. from a Bernoulli distribution of parameter ϕ. Compute an estimate the parameter ϕ using MLE.

Hint: you can write the lihelihood of a single point as $p(x; \phi) = \phi^x (1 - \phi)^{1-x}$.

Solution The likelihood for the n points is:

$$\begin{aligned}\mathcal{L}(\phi) &= \prod_{i=1}^{n} \phi^{x_i}(1-\phi)^{1-x_i} \\ &= \phi^{\sum_{i=1}^{n} x_i}(1-\phi)^{n-\sum_{i=1}^{n} x_i}.\end{aligned}$$

To simplify the notation, let us use the shortcut:

$$S = \sum_{i=1}^{n} x_i,$$

and to simplify the calculations, let us use the log-likelihood:

$$\ell(\phi) = S \log \phi + (n - S) \log(1 - \phi).$$

We find the MLE ϕ^* by solving:

$$\frac{d\ell}{d\phi}(\phi^*) = 0 \Leftrightarrow \frac{S}{\phi^*} - \frac{n - S}{1 - \phi^*} = 0 \Leftrightarrow \phi^* = \frac{S}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

7. Let us consider two boxes. Box 1 contains 30 white balls and 10 black balls. Box 2 contains 20 white and 20 black balls.

   Someone draws one ball at random (among the 80 balls). The drawn ball is white. What is the probability that the ball was drawn from the first box ?

   Solution Let us denote by B_1 (resp. B_2) the event that the ball was drawn from box 1 (resp. 2). Let us denote by W the event that the drawn ball was white.

   We are looking for:

$$\mathbb{P}(B_1 \mid W) = \frac{\mathbb{P}(B_1, W)}{\mathbb{P}(W)}.$$

   We compute:

$$\mathbb{P}(B_1, W) = \mathbb{P}(W \mid B_1)\, \mathbb{P}(B_1) = \frac{3}{4} \times \frac{1}{2} = \frac{3}{8},$$

   and:

$$\mathbb{P}(W) = \frac{50}{80} = \frac{5}{8}.$$

   We conclude:

$$\mathbb{P}(B_1 \mid W) = \frac{\frac{3}{8}}{\frac{5}{8}} = 0.6.$$