1st RECITATION
REVIEW OF PROBABILITY
10-701/15-781 MACHINE LEARNING, FALL 2009
CHARALAMPOS E. TSOURAKAKIS

Abstract.
This is a set of notes and exercises, summarizing and illustrating concepts that we discussed during the
first recitation.

1. Exercises

You can find the definitions that we will use in the following in the book, section 1.2. Here are some exercises
illustrating some important concepts.

Exercise 1 (marginalize, condition, expected value): Let $f_{X,Y}(x,y) = \frac{1}{x}, 0 \leq y \leq x \leq 1$ be the joint pdf of
two random variables $X$ and $Y$.

1. Prove that $f_X(x) = \frac{1}{x}, 0 \leq x \leq 1$

We marginalize: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{0}^{x} \frac{1}{y}dy = \frac{1}{2} \int_{0}^{x}dy = \frac{1}{2}(x - 0) = 1$. Thus, $f_X(x) = 1, 0 \leq x \leq 1$, otherwise $f_X(x) = 0$.

2. Find the conditional density function $f_{Y|X}(y|x)$

By the definition of the conditional density function we have: $f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \frac{1}{x}$ if $0 \leq y \leq x \leq 1$, otherwise 0.

Remarks The answers in 1,2 mean that $X$ is uniformly distributed in $[0, 1]$ and $Y$ given that $\{X = x\}$, $Y$ is uniform
on $[0, x]$.

3. Find the expected values of $Y$

We will use the following important theorem:

Theorem 1. For the discrete case $E(Y) = \sum_x E(Y|X = x)P(X = x)$, and for the continuous case $E(Y) = \int_x E(Y|X = x)f_X(x)$.

Applying the theorem to our problem, given our answers in 1,2: $E(Y) = \int_0^1 E(Y|X = x)f_X(x)dx = \int_0^1 \frac{1}{2}dx = \frac{1}{4}$

Exercise 2 (marginalize, independence): In the class, we said that a discrete random variable can take infinite
values, and I mentioned as an example the Poisson distribution. Consider now two random variables $X$, $Y$ and
you are said that the joint distribution is: $f(x,y) = \frac{\lambda^x}{x!}e^{-\lambda} \frac{\mu^y}{y!}e^{-\mu}$ for $x, y = 0, 1, \ldots$
1. Find the marginal mass function $f_X(x)$.

We use the definition of the marginalization: $f_X(x) = \sum_{y=0}^{\infty} \frac{\lambda^x \mu^y e^{-\lambda + \mu}}{y!}$. Observe that the summation should give 1, since it is the probability mass function of the Poisson distribution. Thus: $f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!}$. By the same argument we get that $f_Y(y) = \frac{\mu^y e^{-\mu}}{y!}$.

Since the discrete random variables $X$ and $Y$ are called independent if and only if $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$ and given our answers, we see that $X$ and $Y$ are two independent Poisson random variables, i.e., $X$ is Poisson($\lambda$), $Y$ is Poisson($\mu$).

**Exercise 3 (law of total probability):** Assume that you have $k$ dollars and you start gambling. The game is as follows: you toss a coin and with probability $1/2$ you win a dollar and with probability $1/2$ you lose one. You play until either you get broke or you win $N$ dollars. What is the probability of getting broke? Even if we are not going to solve the problem entirely we are going to apply the law of total probability to derive a recursive equation.

**Solution** Let $C_k$ be the event of getting broke in the game, having started with $k$ dollars, $W$ the event of winning and $\neg W$ of losing a dollar. By the law of total probability we get $P(C_k) = P(C_k|W) + P(C_k|\neg W) = P(C_k|W)P(W) + P(C_k|\neg W)P(\neg W)$ Now observe that $P(C_k|W) = P(C_{k+1})$ and $P(C_k|\neg W) = P(C_{k-1})$. Thus we got that $P(C_k) = 1/2P(C_{k+1}) + 1/2P(C_{k-1})$.

**Exercise 4 (marginalize, condition, expected value, covariance, correlation):** Consider two random variables $X, Y$ with the following joint probability mass function: $f_{X,Y}(X = 0, Y = 0) = f(0,0) = 0.1$, $f(0,1)=0.3$, $f(1,0)=0.4$ and $f(1,1)=0.2$.

1. **Marginals mass functions of $X, Y$**

   $f_X(X = 0) = f_{X,Y}(X = 0, Y = 0) + f_{X,Y}(X = 0, Y = 1) = 0.1 + 0.3 = 0.4$ and
   
   $f_X(X = 1) = 1 - 0.4 = 0.6$ or $f_X(X = 1) = f_{X,Y}(X = 1, Y = 0) + f_{X,Y}(X = 1, Y = 1) = 0.4 + 0.2 = 0.6$

   In the same way we get: $f_Y(Y = 0) = f_{X,Y}(X = 0, Y = 0) + f_{X,Y}(X = 1, Y = 0) = 0.1 + 0.4 = 0.5$ and
   
   $f_Y(Y = 1) = 1 - 0.5 = 0.5$ or $f_Y(Y = 1) = f_{X,Y}(X = 0, Y = 1) + f_{X,Y}(X = 1, Y = 1) = 0.3 + 0.2 = 0.5$

2. **Expected values of $X, Y$**

   $E(X) = 0 \times f_X(X = 0) + 1 \times f_X(X = 1) = 0.6$ and
   
   $E(Y) = 0 \times f_Y(Y = 0) + 1 \times f_Y(Y = 1) = 0.5$.

3. **Covariance**

   By the definition of the covariance $cov(X, Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$. Thus to compute the covariance of $X, Y$ we have to compute $E(XY)$, i.e. the expected value of the function $g(x, y) = xy$.

   
   $E(XY) = \sum_{x,y} f_{X,Y}(x,y) \times xy = 0 \times 0 \times 0.1 + 0 \times 1 \times 0.3 + 1 \times 0 \times 0.4 + 1 \times 1 \times 0.2 = 0.2$

   Thus the covariance is $cov(X, Y) = E(XY) - E(X)E(Y) = 0.2 - 0.6 \times 0.5 = -0.1$. Thus the two random variables $X, Y$ are correlated.
Exercise 5 (marginalize, conditional mass function, Bayes rule): A factory constructs during a day $N$ lamps, and $N$ follows the Poisson distribution with parameter $\lambda$ which means that $P(N = n) = \frac{\lambda^n}{n!} e^{-\lambda}, n = 0, 1, \ldots$.

Each of the lamps has probability $p$ of being good and $q = 1-p$ of being defective. Let $K$ be the number of lamps in a given day.

1. Compute $E(K|N = n)$

When the number of produced lamps is given, i.e., equal to $n$, then $K$ follows a binomial distribution. Thus, if we attach a simple indicator variable $I_j$ to the $j$-th lamp, i.e., $I_j = 1$ if the lamp works with probability $p$, and $I_j = 0$ if the lamp is defective with probability $1-p$. The expected value of $K$ given $N = n$ is equal to $np$ by the linearity of the expectation. Specifically, $E(K|N = n) = \sum_{j=1}^{n} E(I_j) = \sum_{j=1}^{n} p = np$

Furthermore, observe that the distribution of the random variable $K|N = n$ follows a binomial distribution, i.e., $P(K = k|N = n) = \binom{n}{k} p^k (1-p)^{n-k}$.

2. $E(K)$

Again using theorem 2, we get the following:

$$E(K) = \sum_{n=0}^{\infty} E(K|N = n) P(N = n) = \sum_{n=0}^{\infty} np P(N = n) = p \sum_{n=0}^{\infty} n \times P(N = n) = pE(N) = \lambda$$

since the expected value of a Poisson random variable with parameter $\lambda$ is equal to $\lambda$.

3. Show how to compute the conditional mass function $f(N|K)$

By applying the Bayes rule we get for $n > k$: $f_{N|K}(n|k) = P(N = n|K = k) = \frac{P(K=k|N=n)P(N=n)}{P(K=k)} = \frac{\binom{n}{k} p^k (1-p)^{n-k} \frac{\lambda^n}{n!} e^{-\lambda}}{\sum_{m \geq k} \binom{n}{m} p^m (1-p)^{n-m} \frac{\lambda^m}{m!} e^{-\lambda}} [1]$.

Exercise 6: Disease diagnosis (Bayes rule) Assume that 0.01 % of the population is infected by some disease. Suppose that the test to diagnose if a patient has the disease or not is 99% accurate for both the positive and the negative case, or equivalently the false positive and false negative rate are 0.01. What is the probability of a person that gave a positive test to be sick?

Solution The answer is counterintuitive, even if the test is 99% accurate, a positive result is wrong $\approx 99$% of the time. To see why, define the following events: $P$= the event of a positive test, $D$= the event of having the disease.

We want to compute the probability $Pr(D|P)$, i.e., the probability of a given person being sick given that his/her test was positive. Using the Bayes rule we get the following:

$$Pr(D|P) = \frac{Pr(P|D) Pr(D)}{Pr(P)} = \frac{Pr(P|D) Pr(D)}{Pr(P|D) Pr(D) + Pr(P|\neg D) Pr(\neg D)} = \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.01 \times (1-0.01)} \approx 0.01$$

Exercise 7: To believe them or not? (Bayes rule, chain rule, conditioning) Your friends, John and Mary, who are both fairly reliable told you that they saw a huge comet falling. More specifically, you know that the probability of them telling the truth is 0.95 and 0.9 respectively. i.e., $P(John \text{ tells the truth}) = 0.95$, $P(Mary \text{ tells the truth}) = 0.90$. For brevity, let $T_j = “John \text{ tells the truth}”, T_M = “Mary \text{ tells the truth}”$ and $P(T_j) = 0.95$ and $P(T_M) = 0.9$. Furthermore, the assertions of John and Mary are independent.

\footnote{The result if you carry out the summation is equal to $\frac{(1-p)\lambda e^{-(1-p)\lambda}}{(\lambda-k)!} e^{-(1-p)\lambda}$.}
The probability of the event “a huge comet falls” however is \( \frac{1}{10^6} \). What is the probability that indeed a huge comet has fallen onto earth, given that both John and Mary tell that the event happened?

**Solution** Let C the event that a comet fell, let J and M be the events that John and Mary respectively claim that a comet fell. Observe that \( P(J|C) = P(T_J) = 0.95 \) and \( P(J|-C) = P(-T_J) = 0.05 \) and similarly \( P(M|C) = P(T_M) = 0.9 \) and \( P(M|-C) = P(-T_M) = 0.1 \). In other words, when John says that the comet fell and the comet indeed fell (event \( J \)) then he is telling the truth (event \( T_J \)).

We want to calculate \( P(C|J,M) \). We will use Bayes rule:

\[
P(C|J,M) = \frac{P(C,J,M)}{P(J,M)} = \frac{P(J|C)P(M|C)P(C)}{P(J,M)} = \frac{P(J|C)(M|C)P(C)}{P(J,M)} = \frac{P(J|C)(M|C)P(C)}{P(J)(M|C)+P(J|-C)(M|-C)P(-C)}
\]

\[
= \frac{0.95 \times 0.9 \times 10^{-6}}{0.95 \times 0.9 \times 10^{-6} + 0.95 \times 0.1 \times 10^{-6}} \approx 1.71 \times 10^{-4}
\]

2. Least Squares

I mentioned few things about the least squares approach and the geometric intuition behind it, just to help you with one question in your homework.

2.1. **Projections.** Consider a vector \( b \) and a vector \( a \). You are asked to find the projection \( p \) of \( b \) on \( a \) and furthermore write that projection as \( p = Pb \), i.e., write the projection as a matrix vector multiplication.

Consider the vector \( e = b - p \). This vector by the definition of the projection is orthogonal to the vector \( a \), which gives the following equation:

\[
a^T e = 0 \rightarrow a^T (b - p) = 0 \rightarrow a^T b = a^T p
\]

Notice that \( p = xa \) for some scalar \( x \). Substituting to the above equation gives the following \( a^T xa = a^T b \rightarrow x = \frac{a^T b}{a^T a} \). Thus, the projection \( p \) we wanted is \( p = xa = a^T b/a^T a \).

Before answering the second question (find the matrix \( P \) s.t \( Pb = p \)) few remarks on this result. What happens when you scale \( a \) for example you consider the projection of \( b \) on \( 2a \)? The answer of course is that the projection \( p \) should not change. This comes also from the formula, you get a 4 in the nominator and another 4 in the denominator. What happens when \( b \) and \( a \) are of the form \( b = \lambda a \), for some scalar \( \lambda \)? Let’s substitute:

\[
p = a^T x a \rightarrow p = a^T \lambda a = \lambda b
\]

In that case, the projection of \( b \) is \( b \) again. Now consider the case of \( b \) being orthogonal to \( a \). We see that the projection is 0. Now let’s see again the same formula: \( p = a^T x a \rightarrow p = (a^T x a) b = Pb \). Therefore the matrix \( P \) projects vector \( b \) on the \( a \).

Why do we care about projections in this class? Consider the following example, you have a black box that takes input \( x \) and outputs a \( y \), equal to \( y = c_1 x + c_0 \). If we had no noise, we could just try two different \( x \)’s, e.g., \( x_1, x_2 \), measure two corresponding \( y \)’s, \( y_1, y_2 \) and solve a simple linear system with two equations and two unknowns. Assume however that when you measure \( y \) with some device, a small amount of error is introduced, and that you have \( n \) measurements, where \( n \) is much larger than 2. You can express this as a linear system: \( y = Xc \) where \( c = [c_0 c_1]^T \), \( y \) is the vector containing the outputs/measurements \( y = [y_1 \ldots y_n]^T \) and \( X \) is a \( n \times 2 \) matrix that contains in the \( i \)-th row \( x_{i1} = 1, x_{i2} = x_i \), i.e., in the first column the constant 1 and in the second the input value
Figure 1. Blackbox

Figure 2. Any vector of the form $Xc$ where $X$ has two columns $x_1, x_2$ lies on the subspace defined by them. Thus if we have a system to solve of the form $Xc = y$ and $y$ does not lie on that space the system will have no solution. This is the case where we want to project $y$ on that space. In the following lectures and recitations we will discuss more on this problem.

$x_i$ (to understand why, just do the matrix vector multiplication). Let’s look a little bit more the right hand side $Xc$. As we discussed this vector is a linear combination of the two columns of $X$, call them $x_1, x_2$, i.e. $X = [x_1| x_2]$. $Xc = c_0 \times x_1 + c_1 \times x_2$. Thus it lies on the subspace defined by those two vectors. Chances are that $y$ the left handside of your system does not lie on that subspace and thus the system does not have a solution. Then, we said is that an optimal thing to do (with respect to the least squares approach) is to project $y$ on that subspace and solve a system which has a solution. More about the least squares approach will follow in a next recitation, after we will have discussed it in the class. The above are shown in figure 2.

---

\(^2\) Note however, if our assumption that errors are small holds, it should lie “close” to that subspace.