CMPT 882 Machine Learning

Lecture Notes

Instructor: Dr. Oliver Schulte

Scribe: Qidan Cheng and Yan Long

Mar. 9, 2004 and Mar. 11, 2004

Basic Definitions and Facts from Statistics

1. The Binomial Distribution

Given a worn and bent coin, there are two possible outcomes of the coin, head and tail, when tossed. Let the probability of turning up heads be p, and toss the coin n times:

Question: In n independent tosses (trials), what's the probability of observing exactly r heads?

1.1 What's the Binomial distribution?

Binomial distribution describes for each possible value of r ($0 \le r \le n$), the probability of observing exactly r heads given a sample of n independent tosses of a coin whose true probability of heads is p.

In order to derive the binomial probability, let us look at a simple coin toss example, where the probability of turning up head is 0.5 (fair coin).

Assume $n = 3$, there are eight possible outcome sequences:

Trial 1	Trial 2	Trial 3	# of heads
T	T	T	0
T	T	H	1
T	H	T	1
T	H	H	2
H	T	T	1
H	T	H	2
H	H	T	2
H	H	H	3

The probability of each of these sequences is the same. Since there are eight sequences, the probability of each sequence is 1/8. So the probability of each number of heads can be computed:

# of heads	Sequences	probability
0	1	1/8
1	3	3/8
2	3	3/8
3	1	1/8

Now, we wish to generalize the above probability when the probability that coin comes head is p. Since p is the probability of being a head, then $1-p$ is the probability that the coin comes up tail.

For the case of three flips of a coin whose probability of coming up heads is p, we have:

Trial 1	Trial 2	Trial 3	# of heads	probability
T	T	T	0	$(1-p)\,(1-p)\,(1-p)$
T	T	H	1	$(1-p)\,(1-p)\,p$
T	H	T	1	$(1-p)\,p\,(1-p)$
T	H	H	2	$(1-p)\,p\,p$
H	T	T	1	$p\,(1-p)\,(1-p)$
H	T	H	2	$p\,(1-p)\,p$
H	H	T	2	$p\,p\,(1-p)$
H	H	H	3	$p\,p\,p$

Note that for each sequence, the probability of the sequence σ occurs is $p^r(1-p)^{n-r}$, where r is the number of heads in this sequence. And the probability of observing exactly r heads in n tosses is:

$$\Pr(\text{ } r \text{ success in } n \text{ trials}) = \sum_{r \text{ heads } |\sigma|=n} p^r(1-p)^{n-r}$$

Question: How many sequences contain exactly r heads in n trials?

1.2 Permutation and Combination

Permutation:
A permutation of n objects is an ordering of the n objects. An **r-permutation** of n objects is an ordering of an r-element subset of the n objects.

Sampling without replacement: you don't replace the item selected in the population after each draw, so each outcome depends on all previous outcomes.

Example:
Without replacement, how many different ways could we draw 3 samples from elements A, B, C?

Solution: 1. Choose the first letter (3 choices),
2. Choose the second letter (2 choices),
3. Choose the third letter (1 choice)

Number of possible ways is 3*2*1 = 6.

This is exactly the number of permutations of 3 objects.

In general, let $P_{n,n}$ denotes the number of n-permutations of n objects, then

$$P_{n,n} = n*(n-1)*\ldots*1 = n!$$

Example:
Given 25 students, we want to form a club committee consisting of a president and a secretary, how many ways can they select the committee (which is equivalent to drawing two samples from 25 elements without replacement)?

Solution: 1. Choose the president (25 choice)
2. Choose the secretary (24 choices)

Number of possible ways is 25*24=600

This is exactly the number of distinct 2-permutations of 25 objects.

In general, let $P_{n,r}$ denotes the number of r-permutations of n objects, then

$$P_{n,r} = n\,(n-1)\ldots(n-r+1) = n!\,/\,(n-r)!$$

Combination:
A selection of a finite number (n) of objects without regard to order is called a combination. An **r-combination** is an unordered selection of r elements from a set of n objects. We denote $C_{n,r}$ as the number of r-combinations of n objects.

Question: How is the number of combinations $C_{n,r}$ relate to the number of permutations $P_{n,r}$?

Example:
Consider an example of $C_{4,3}$, all 3-combinations of $\{a, b, c, d\}$ are:

abc abd acd bcd

Each of the combination above will give rise to 3! Permutations

abc abd acd bcd
acb adb adc bdc
bac bad cad cbd
bca bda cda cdb
cab dab dac dbc
cba dba dca dcb

Each *column* is the 3! permutations of that combination. But those permutations are the same combination – because the order does not matter. Therefore, there are 3! times as many permutations as combinations. $C_{4,3}$, therefore, will be $P_{4,3}$ divided by 3! – the number of *permutations* that each combination generates.

$$C_{4,3} = P_{4,3} / 3!$$

In general, the way of constructing $P_{n,r}$ from $C_{n,r}$:

- Choose a member s of $C_{n,r}$ (with r elements)
- Output all permutations of $s \rightarrow P_{r,r} = r!$

$$P_{n,r} = C_{n,r} \cdot r!$$
$$C_{n,r} = P_{n,r} / r! = n! / ((n-r)! \, r!)$$

Now we can come back to our original question: How many sequences contain exactly r heads in n trials? This question is equivalent to: what is the number of ways to pick which r of the n trials are heads. The answer is $C_{n,r}$.

Thus, we have the **binomial distribution**:

$$\Pr(r \text{ success in } n \text{ trials}) = \sum_{r \text{ heads } |\sigma|=n} p^r (1-p)^{n-r}$$
$$= C_{n,r} \, p^r (1-p)^{n-r}$$
$$= (n! / ((n-r)! \, r!)) \, p^r (1-p)^{n-r}$$

2. Random Variables

2.1 Definitions and Notations:

A **random variable** (r.v.) X is a pair $X = <D_x, P_x)$ where D_x is the domain of X and P_x is the probability distribution over Dx.

Probability distribution is defined for every number x by $p(x)$ or $P(X = x)$. It is also called probability mass function. In words, for every possible value x of the random variable, probability distribution specifies the probability of observing that value when the experiment is performed.

Notations:
X(capital letters) = the random variable.
x (small letters) = a number that the discrete random variable could assume.
$P(X = x)$: is the probability that the random variable X equals x. $P(X = x)$ can also be represented by $p(x)$.

Example:
For our coin toss example, we can define random variable C = < {H,T}, { p(H) = 1/2, p(T) = 1/2} >. {H,T} is the domain of random variable C, the probability of turning up head is 1/2, and the probability of turning up tail is also 1/2.

2.2 Scalar operation over random variable

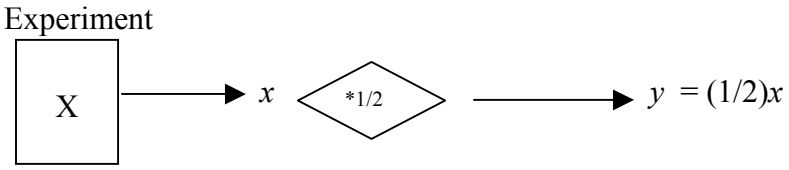
When the domain of a random variable is real number, we can generate new random variable over the original one. For example we can define scalar operation over this random variable.

Example:

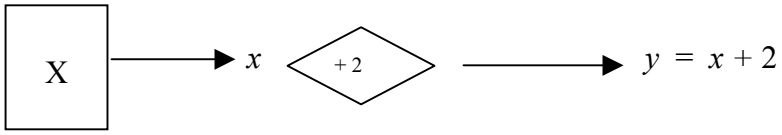
$X = <\{0,1\}, P_X>$ where $P_X(1) = 1/2$, then what does aX mean? Clearly, aX is another random variable. So, we can define this new random variable as follows:

$Y = [aX] = <R, P_Y>$, where $P_Y(y) = P_X\{x: ax = y\} = \sum_{x: ax = y} P_X(x)$

Note: $y = ax$ is an one-to-one mapping. In general, we may have many-to-one mapping, so we need to sum all the x's which are mapped to the same y.



$P_Y(1/2) = P_X(1) = 1/2$ $\quad$ $P_Y(0) = P_X(0) = 1/2$



$P_Y(3) = P_X(1) = 1/2$ $\quad$ $P_Y(2) = P_X(0) = 1/2$

2.3 Joint Probability Distribution and Marginal probability

Let X and Y be two random variables defined on the sample space of an experiment. The joint probability distribution $p(x, y)$ is defined over X, Y for each pair of numbers (x, y) by:

$$p(x, y) = P(X = x \text{ and } Y = y) = P(X = x, Y = y)$$

- In general, P_X and P_Y do not give us joint probability $P(X = x, Y = y)$
- If x, y is independent, then $P(X = x, Y = y) = P_X(x) * P_Y(y)$
- From the additivity and normalization axioms of probability, $\sum_{x, y} p(x, y) = 1$

Marginal probability: The individual probability distribution of a random variable is referred to as its marginal probability distribution.

Lemma: Given two random variables X = <Dx, Px> and Y= <D_Y, P_Y> with joint probability distribution p(x, y), we can determine their marginal probability distribution P(X = x) and P(Y = y).

$$P(X = x) = \sum_{y \in P_Y} p(x, y) \qquad P(Y = y) = \sum_{x \in P_X} p(x, y)$$

Proof: P(X = x) is logically equivalent to

$$\begin{cases} P(X = x \wedge Y = y_0) \text{ or} \\ P(X = x \wedge Y = y_1) \text{ or} \\ P(X = x \wedge Y = y_2) \text{ or} \\ \quad . \\ \quad . \\ \quad . \\ P(X = x \wedge Y = y_n) \end{cases}$$

$$\Rightarrow \qquad P(X = x) = \sum_{y \in P_Y} p(x, y)$$

Similarly, we can prove $P(Y = y) = \sum_{x \in P_X} p(x, y)$

Note:

- We can derive the marginal probabilities of X and Y from the joint probability over X and Y, but we can not get joint probability over X and Y from the marginal probabilities of X and Y.
- From the marginal probabilities of X and Y, the probabilities of events involving only X or only Y can be computed.

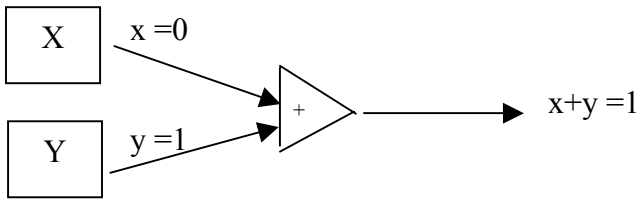
Example: Given the joint probability over X and Y, what is the marginal probability of X and Y?

X \ Y	0	1	Marginal probability of X
0	0.3	0.2	0.5
1	0.25	0.25	0.5
Marginal probability of Y	0.55	0.45	

2.4 Sum of two random variables

Given two random variables X and Y, and the joint probability distribution p(x, y) over X,Y, we can define Z = X + Y as follows:

$Z = \langle R, P_Z \rangle$ where $P(z) = \sum_{x \in Dx,\, y \in D_Y : x+y=z} p(x, y)$



Example:

Given X, Y are two random variables representing coin toss example, and let value 1 denote the outcome head, and 0 denote the outcome tail. The joint probability distribution over X and Y is as follows, define Z = X + Y:

X \ Y	0	1
0	1/4	1/4
1	1/4	1/4

$P_Z(Z=0) = P(X= 0, Y = 0) = 1/4$

$P_Z(Z=1) = P(X = 0, Y = 1) + P(X=1, Y=0) = 1/4 + 1/4 = 1/2$

$P_Z(Z=2) = P(X = 1, Y = 1) = 1/4$

2.5 Expected Value of a random variable

The **expected value**, or **mean**, of a random variable X is defined by

$E[X] = \sum_{x \in Dx} x\, p(x)$ for $Dx \subseteq R$

($E[X] = \int x\, p(x)dx$ for a continuous random variable X)

Note: E [X] may not be defined. E [X] is well defined if the sum is finite or converges absolutely. Sometimes the expected value of X is denoted by μ_X or, when the random variable is apparent from the context, simply by μ.

Example:

Let H_1 and H_2 be two random variables, and
$H_1 = < \{0, 1\},\ \{p(1) = 1/2,\ p(0) = 1/2\} >$
$H_2 = < \{0, 1\},\ \{p(1) = 1/2,\ p(0) = 1/2\} >$

$E[H_1] = 0 \times 1/2 + 1 \times 1/2 = 1/2$
$E[H_2] = 0 \times 1/2 + 1 \times 1/2 = 1/2$

H_1 \ H_2	1	0
1	2, 1/4	1, 1/4
0	1, 1/4	0, 1/4

$H_1 + H_2 = < \{0, 1, 2\},\ \{p(0) = 1/4,\ p(1) = 1/2,\ p(2) = 1/4\} >$

$E[\ H_1 + H_2\ ] = 0 \times 1/4 + 1 \times 1/2 + 2 \times 1/4 = 1$

Linearity of Expected Value of Random Variable

Theorem 1: Let X be a random variable, $Y = aX + b$, where a and b are any constants. If E[X] is well defined, then

$$E[\ Y\ ] = E[\ aX + b\ ] = a \cdot E[\ X\ ] + b$$

Proof:

$$E[\ Y\ ] = E[\ aX + b\ ]$$

$$= \sum_{x \in Dx} (ax + b) \cdot p(x)$$

$$= \sum_{x \in Dx} [ax \cdot p(x) + b \cdot p(x)]$$

$$= \sum_{x \in Dx} ax \cdot p(x) + \sum_{x \in Dx} b \cdot p(x)$$

$$= a \cdot \sum_{x \in Dx} x \cdot p(x) + b \cdot \sum_{x \in Dx} p(x)$$

$$= a \cdot E[X] + b \qquad \left(\sum_{x \in Dx} p(x) = 1\right)$$

Theorem 2: Let X and Y be random variables, and $p(x, y)$ be the joint probability distribution of X and Y. Assume E [X] and E [Y] are well defined, then

$$E[X + Y] = E[X] + E[Y]$$

Proof:

$$E[X + Y] = \sum_{x \in Dx} \sum_{y \in Dy} (x + y)\, p(x, y)$$

$$= \sum_{x \in Dx} \sum_{y \in Dy} [\, x \cdot p(x, y) + y \cdot p(x, y)\,]$$

$$= \sum_{x \in Dx} \sum_{y \in Dy} x \cdot p(x, y) + \sum_{x \in Dx} \sum_{y \in Dy} y \cdot p(x, y)$$

$$= \sum_{x \in Dx} \sum_{y \in Dy} x \cdot p(x, y) + \sum_{y \in Dy} \sum_{x \in Dx} y \cdot p(x, y)$$

$$= \sum_{x \in Dx} x \sum_{y \in Dy} p(x, y) + \sum_{y \in Dy} y \sum_{x \in Dx} p(x, y)$$

$$= \sum_{x \in Dx} x\, p(x) + \sum_{y \in Dy} y\, p(y) \qquad \left(\sum_{y \in Dy} p(x, y) = p(x), \sum_{x \in Dx} p(x, y) = p(y) \right)$$

$$= E[X] + E[Y]$$

Note: This theorem holds no matter X and Y are independent or not.

Theorem 3: Let X and Y be two independent random variables, and $p(x, y)$ be the joint probability distribution of X and Y. Assume E [X] and E [Y] are well defined, then

$$E[X\,Y] = E[X] \cdot E[Y]$$

Proof:

$$E[X\,Y] = \sum_{x \in Dx} \sum_{y \in Dy} (x\, y)\, p(x, y)$$

$$= \sum_{x \in Dx} \sum_{y \in Dy} (x\, y)\, p(x)\, p(y)$$

$$= \sum_{x \in Dx} \sum_{y \in Dy} (\, x\, p(x)\,)\,(\, y\, p(y)\,)$$

$$= \sum_{x \in Dx} x\, p(x) \sum_{y \in Dy} y\, p(y)$$

$$= E[X]\, E[Y]$$

Expected Value of the Sample Mean

The random variables $H_1, H_2, \ldots, H_n$ are said to form a **random sample** of size n if

1. The H_i's are independent random variables.
2. Every H_i has the same probability distribution.

Let H be a random variable defined over the population, $H_1, H_2, \ldots, H_n$ be a random sample from the distribution with mean value E [H], $\overline{H} = (H_1 + H_2 + \ldots + H_n) / n$ be the **sample mean**, then

$$E [H_1 + H_2 + \ldots + H_n] = n \cdot E [H]$$

$$\begin{aligned} E [\overline{H}] &= E [(H_1 + H_2 + \ldots + H_n) / n] \\ &= (1/n) E [H_1 + H_2 + \ldots + H_n] \\ &= (1/n) \cdot n \cdot E [H] \\ &= E [H] \end{aligned}$$

According to the above result, the sampling (i.e., probability) distribution of the sample mean $\overline{H}$ is centered precisely at the mean of the population from which the sample has been selected. The important of the sample mean $\overline{H}$ springs from its use in drawing conclusions about the population mean.

2.6 The Variance of Random Variable

The expected value of a random variable does not tell us how "spread out" the variable's value are. For example, suppose we have random variables X and Y,

$X = < \{2\}, \{ P(2) = 1\} >$
$Y = < \{ 1, 2, 3\}, \{ P(1) = P(2) = P(3) = 1/3 \} >$

$E [X] = 2 \times 1 = 2$
$E [Y] = 1 \times 1/3 + 2 \times 1/3 + 3 \times 1/3 = 2$

	X	Y
3		*
2	* * *	*
1		*

Both E[X] and E[Y] are 2, yet the actual values taken on by Y are farther from the mean than the actual values taken on by X.

Definition: Let X be a random variable with expected value E [X] = μ, then the **variance** of X, denoted by Var[X], is

$$\text{Var}[X] = E[(X - E[X])^2] = E[(X - \mu)^2]$$

Example: Suppose we have a random variables $X = \langle \{0, 1\}, \{P(1) = p, P(0) = 1 - p\} \rangle$, then

$$\mu = E[X] = 1 \times p + 0 \times (1 - p) = p$$

$$\begin{aligned}
\text{Var}[X] &= E[(X - \mu)^2] \\
&= \sum_{x \in Dx} (x - \mu)^2 \; p(x) \\
&= (1 - p)^2 \cdot p(1) + (0 - p)^2 \cdot p(0) \\
&= (1 - p)^2 \cdot p + p^2 \cdot (1 - p) \\
&= p(1 - p)(1 - p + p) \\
&= p(1 - p)
\end{aligned}$$

Lemma: Let X be a random variable, then Var[X] = 0 if and only if there exists a constant *c* such P(X = *c*) = 1.

Theorem 3: Let X be a random variable, *a* and *b* be any constants, then

$$\text{Var}[aX + b] = a^2 \cdot \text{Var}[X]$$

Proof: Let E [X] = μ

$$\begin{aligned}
\text{Var}[aX + b] &= E[((aX + b) - E[aX + b])^2] \\
&= E[((aX + b) - (a \cdot E[X] + b))^2] \qquad \text{(By Theorem 1)} \\
&= E[((aX + b) - (a\mu + b))^2] \\
&= E[(aX - a\mu)^2] \\
&= E[a^2 (X - \mu)^2] \\
&= a^2 \cdot E[(X - \mu)^2] \\
&= a^2 \cdot \text{Var}[X]
\end{aligned}$$

Theorem 4: Let X and Y be independent random variables, then

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$$

Proof:

Let $E[X] = \mu_x$, $E[Y] = \mu_y$

Then $E[X + Y] = E[X] + E[Y] = \mu_x + \mu_y$

$$\text{Var}[X + Y] = E[((X + Y) - E[X + Y])^2]$$

$$= E[((X + Y) - (\mu_x + \mu_y))^2]$$

$$= E[((X - \mu_x) + (Y - \mu_y))^2]$$

$$= E[(X - \mu_x)^2 + 2(X - \mu_x)(Y - \mu_y) + (Y - \mu_x)^2]$$

$$= E[(X - \mu_x)^2] + 2E[(X - \mu_x)(Y - \mu_y)] + E[(Y - \mu_x)^2]$$

$$= \text{Var}[X] + 2E[(X - \mu_x)(Y - \mu_y)] + \text{Var}[Y]$$

$$E[(X - \mu_x)(Y - \mu_y)] = \sum_{x \in Dx} \sum_{y \in Dy} (x - \mu_x)(y - \mu_y)\, p(x, y)$$

$$= \sum_{x \in Dx} \sum_{y \in Dy} (x - \mu_x)(y - \mu_y)\, p(x)\, p(y)$$

$$= \sum_{x \in Dx} (x - \mu_x)\, p(x) \sum_{y \in Dy} (y - \mu_y)\, p(y)$$

$$= E[(X - \mu_x)]\, E[(Y - \mu_y)]$$

$$= (E[X] - E[\mu_x])(E[Y] - E[\mu_y])$$

$$= (\mu_x - \mu_x)(\mu_y - \mu_y)$$

$$= 0$$

Hence,

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$$

More about Variance

$\text{Var}[aX + b] = a^2 \cdot \text{Var}[X]$
(This result says that the addition of the constant b does not affect the variance, which is intuitive, because the addition of b changes the location (mean value) but not the spread of the value)

$\text{Var}[X] = \text{Var}[-X]$

Variance of the Sample Mean

Let H be a random variable defined over the population, $H_1, H_2, \ldots, H_n$ be a random sample from the distribution with variance Var [H], $\overline{H} = (H_1 + H_2 + \ldots + H_n) / n$ be the sample mean, then

$$\text{Var}[H_1 + H_2 + \ldots + H_n] = n \cdot \text{Var}[H]$$

$$\begin{aligned}\text{Var}[\overline{H}] &= \text{Var}[(H_1 + H_2 + \ldots + H_n) / n] \\ &= (1/n^2) \cdot \text{Var}[H_1 + H_2 + \ldots + H_n] \\ &= (1/n^2) \cdot n \cdot \text{Var}[H] \\ &= (1/n) \cdot \text{Var}[H]\end{aligned}$$

According to the above result, the sample mean $\overline{H}$ distribution becomes more concentrated as the sample size n increases.

References:

1. Notes taken from Dr. Schulte's class on the above-mentioned dates.
2. Probability and statistics for engineering and the sciences 4th edition, Jay L. Devore, Belmont: Duxbury Press, c1995