Consider the problem of predicting one's risk of cancer based on the average number of cigarettes smoked per day.

Let X be a random variable that is the average number of cigarettes a person smokes per day. Let C1 be the class of high risk of lung cancer and C2 be the one with low risk.

We use Bayesian decision theory to solve this problem. This means we have to determine P(C1|X) and P(C2|X) for a given X and choose the class with the higher probability. According to Bayes rule we have

$P(C1|X) = P(X|C1)P(C1) / P(X)$
(posterior) = (likelihood)*(prior)/(normalization)

and similarly for P(C2|X).

What is P(X)? $P(X) = P(X|C1)P(C1) + P(X|C2)P(C2)$

Suppose we randomly sampled 1100 people from the population and suppose 100 of them have lung cancer and 1000 are controls. We asked each person the average number of cigarettes they smoked over some previous time period and obtained the following data.

Average number of cigarettes smoked per day	Between 0 and 4	Between 5 and 9	Between 10 and 14	Between 15 and 19
C1	5	20	35	40
C2	900	80	15	5

Estimate probabilities P(C1), P(C2), P(X between a and b|C1) (for the four ranges of a and b above), and P(X between a and b|C2) from the table data above.

Solution:

Why is P(C1) = Number of people in class C1/(total samples) ?
What is P(C1) conceptually? It is the probability that a person has lung cancer.

Assumptions:

Each individual is chosen independently and identically (i.i.d).

To estimate P(C1) imagine a coin-tossing analogy

P(C1) = 100/1100
P(C2) = 1000/1100

P(X between 0 and 4|C1) = 5/100 = .05
P(X between 5 and 9|C1) = 20/100 = .2
P(X between 10 and 14|C1) = .35
P(X between 15 and 19|C1) = .4

P(X between 0 and 4|C2) = 900/1000 = .9
P(X between 5 and 9|C2) = .08
P(X between 10 and 14|C2) = .015
P(X between 15 and 19|C2) = .005

Now suppose you encounter an individual who smokes an average of 2 cigarettes per day. Are they at high risk for lung cancer?

You need to calculate P(C1|X) and P(C2|X) and pick the higher probability to make your decision.

P(C1|X=2) = P(X=2|C1)P(C1)/P(X)
= (.05 * 100/1100)/P(X) = .0045/P(X)

P(C2|X=2) = P(X=2|C2)P(C2)/P(X)
= (.9 * 1000/1100)/P(X) = 0.818/P(X)

P(C1|X=2) + P(C2|X=2) should sum to 1. Does it? Let's check.

.0045/P(X) + .818/P(X) = (.0045+.818)/P(X) = .823/.823 = 1

P(X) = P(X|C1)P(C1) + P(X|C2)P(C2) = .0045 + .818 = 0.823

Since P(C2|X=2) > P(C1|X=2) person X is at low risk for lung cancer

But what if someone smoked an average of 10 per day?

P(C1|X=10) = P(X=10|C1)P(C1)
= .35 * 100/1100 = 35/1100

P(C2|X=10) = .015 * 1000/1100 = 15/1100

So if someone smoked an average of 10 per day they would be at high risk for lung cancer

Let's move on. In practice you may have more than one variable to create a decision rule. For example for our current problem, suppose we add average number of

hours a person exercises.

X = average number of cigarettes a person smokes per day
Y = average number of hours a person exercises per day

Bayesian decision theory dictates we calculate

P(C1|X,Y) and P(C2|X,Y)

P(C1|X,Y) = P(X,Y|C1)P(C1)

If we assumed that X and Y are independent then what is P(X,Y|C1)?

P(X,Y|C1) = P(X|C1)P(Y|C1)

We already have P(X|C1) and P(X|C2)

P(X between 0 and 4|C1) = 5/100 = .05
P(X between 5 and 9|C1) = 20/100 = .2
P(X between 10 and 14|C1) = .35
P(X between 15 and 19|C1) = .4

P(X between 0 and 4|C2) = 900/1000 = .9
P(X between 5 and 9|C2) = .08
P(X between 10 and 14|C2) = .015
P(X between 15 and 19|C2) = .005

Suppose someone gave us P(Y|C1) as well

P(Y between 0 and .25|C1) = .8
P(Y between .25 and .5|C1) = .15
P(Y between .5 and .75|C1) = .04
P(Y between .75 and 1|C1) = .01

P(Y between 0 and .25|C2) = .5
P(Y between .25 and .5|C2) = .35
P(Y between .5 and .75|C2) = .1
P(Y between .75 and 1|C2) = .05

Suppose we added a third variable Z = age of person. So now P(X,Y,Z|C1) = P(X|C1)P(Y|C1)P(Z|C1). How do we determine P(Z|C1)? We need to collect data. Suppose we collected the data into the table below

Age of person	Between 0 and 15	Between 15 and 30	Between 30 and 45	Above 45
C1	1	3	30	66
C2	100	600	250	50

From the above table we can estimate the probabilities that we need.

Let's assume that X and Y are not independent.

$P(C1|X,Y) = P(X,Y|C1)P(C1)$

How can we estimate these probabilities? Consider the table shown below:

Average number of cigarettes smoked per day and average number of hours exercised per day is between 0 and 0.25	Between 0 and 4	Between 5 and 9	Between 10 and 14	Between 15 and 19
C1	5	20	35	40
C2	900	80	15	5

Average number of cigarettes smoked per day and average number of hours exercised per day is between 0.25 and 0.5	Between 0 and 4	Between 5 and 9	Between 10 and 14	Between 15 and 19
C1	3	15	20	25
C2	917	70	10	3

Average number of cigarettes smoked per day and average number of hours exercised per day is between 0.5 and 0.75	Between 0 and 4	Between 5 and 9	Between 10 and 14	Between 15 and 19
C1	2	10	17	20
C2	950	50	5	2

Average number of cigarettes smoked per day and average number of hours exercised per day is between 0.75 and 1	Between 0 and 4	Between 5 and 9	Between 10 and 14	Between 15 and 19
C1	1	2	10	30
C2	960	40	5	2

P(C1) = 255/(255+1007+1007+1000+1000) = 255/(4014+255) = 0.0597

P(X is between 5 and 9, Y between 0.5 and 0.75|C1) = 10/255 = .039

What is the above probability if X and Y are independent?

P(X is between 5 and 9, Y between 0.5 and 0.75|C1)=
P(X is between 5 and 9|C1) * P(Y between 0.5 and 0.75|C1)=

What is P(X between 5 and 9|C1) = 47/255
What is P(Y between 0.5 and 0.75)|C1) = 49/255

And so the answer changes if we assume independence and becomes 47*49/(255*255).

If we have two variables X and Y where each variable has say four categories (or intervals).

P(X between some interval and Y between some interval)

Since we have four intervals for each X and Y we have 4 * 4 = 16 probabilities to estimate.

Now suppose we have another variable Z that also has four categories. How many total probabilities do we have to estimate? 4 * 4 * 4 = 64.

Suppose we have 10 variables each with four categories. How many total probabilities do we have to estimate? 4^10 = 1,048,576

If we have 20 variables then how many total probabilities? 4^20 = 1,099,511,627,776 or approximately 10^12.

Going to one variable, what if we assumed that P(X|C1) follows a Gaussian distribution.

Recall when we started the module we talked about coin tosses. Specifically we talked about tossing a coin n times and counting the number of n trials where we have k heads. We noticed that the P(heads=k) was given by a binomial distribution which becomes Gaussian as n approaches infinity or as you get more samples.

P(X=k|C1) = Gaussian(m1, sd1)

P(X=k|C2) = Gaussian(m2,sd2)

Are we ready to do some prediction?

What is P(C1|X=1.5)? We know that P(C1|X=1.5) = P(X=1.5|C1)P(C1)

Suppose P(C1)=.1 and P(C2)=.9. What is P(X=1.5|C1)? We assumed it is Gaussian with mean m1 and standard deviation sd1. Therefore

P(X=1.5|C1) = Gaussian(X=1.5,mean=m1,sd=sd1)

We can't use the above likelihoods in our Bayesian decision rule because we don't know m1, m2, sd1, and sd2. We need these values to calculate

P(C1|X=k) and P(C2|X=k)

We need data to estimate m1, m2, sd1, and sd2, just like we used data to calculate the likelihoods directly (see above). We collect data as we did previously. Assume we have the data below (as an example)

Suppose we sampled 10 people who have lung cancer and recorded their mean number of cigarettes smoked per day:

C1: 10, 5, 8, 15, 17, 18, 2, 3, 11, 12 (training data)

Xi represents the average number of cigarettes smoked per day by individual i.

Pr(x1=10, x2=5, x3=8, x4=15, x5=17, x6=18, x7=2, x8=3, x9=11, x10=12|C1)= Pr(x1=10|C1)Pr(x2=5|C1)Pr(x3=8|C1)Pr(x4=15|C1)Pr(x5=17|C1)Pr(x6=18|C1)Pr(x7=2|C1)Pr(x8=3|C1)Pr(x9=11|C1)Pr(x10=12|C1)

Suppose we sampled 10 people who are controls (don't have lung cancer) and recorded their mean number of cigarettes smoked per day:

C2: 2, 0, 0, 1, 10, 0, 2, 4, 1, 0 (training data)

Unknown: 10, 2, 15 (test data)

We want to find Gaussian parameters that maximize the likelihood. This means we have to find m1 and sd1 that maximizes Pr(x1=10|C1)Pr(x2=5|C1)Pr(x3=8|C1)Pr(x4=15|C1)Pr(x5=17|C1)Pr(x6=18|C1)Pr(x7=2|C1)Pr(x8=3|C1)Pr(x9=11|C1)Pr(x10=12|C1)

Suppose we have f(x). How to find x that maximizes f(x)?

We find df/dx and solve for df/dx = 0. This works mainly if there is only one unique global maximum.

The maximum likelihood method tells us that the mean and variance that maximize the likelihood are just the sample mean and variance.

Let log be the natural log (also denoted as ln)
Recall that

1. ln(xy) = ln(x) + ln(y)
2. ln(x)^y = y ln(x)
3. ln(x/y) = ln(x) – ln(y)
4. ln(e) = 1
5. ln(a * e^x) = ln(a) + ln(e)^x = ln(a) + x*ln(e) = ln(a) + x

Suppose we find x* such that ln(f(x*)) is maximum over all x.

Question: Is x* also the maximum of f(x)?

Yes because ln(x) is monotonically increasing

So now returning to our problem.

Suppose we sampled 10 people who have lung cancer and recorded their mean number of cigarettes smoked per day:

C1: 10, 5, 8, 15, 17, 18, 2, 3, 11, 12

Xi represents the average number of cigarettes smoked per day by individual i.

We assume that P(X=k|C1) is Gaussian distributed. What is the mean and variance of that Gaussian distribution?

So our Gaussian mean m1 = 10.1 and the variance v1 = 28.97 (sd1 = 5.38)

Suppose we sampled 10 people who are controls (don't have lung cancer) and recorded their mean number of cigarettes smoked per day:

C2: 2, 0, 0, 1, 10, 0, 2, 4, 1, 0

We assume here also that P(X=k|C2) is Gaussian with mean m2 and variance v2.

What are the maximum likelihood estimates of the mean and variance?

m2 = 2 v2 = (0 + 4 + 4 + 1 + 64 + 4 + 0 + 4 + 1 + 0) / 10 = 8.2 sd2 = 2.86

Now we are ready to make decisions:

P(C1) = .06
P(C2) = 1-.06 = .94

Suppose we have an individual who smokes average of 5 cigarettes per day. To determine if they are at high risk of lung cancer we calculate

P(C1|X=5) and P(C2|X=5)

P(C1|X=5) = P(X=5|C1)P(C1) = Gaussian(X=5,m1=10.1, sd1=5.38)(0.06)

P(C2|X=5) = P(X=5|C2)P(C2) = Gaussian(X=5,m1=2, sd1=2.96)(0.94)

Now suppose we have two random variables X (mean number of cigarettes smoked per day) and Y (mean number of hours exercised per day).

Let x = (X, Y)

We assume that P(X,Y|C1) is Gaussian distributed. What does the multivariate Gaussian distribution look like?

Let's do multivariate for two variables X and Y. Our mean is a vector of two dimensions and what is Σ ? Σ is a matrix of dimension 2 by 2.

Σ = (Var(X) Cov(X,Y)
(Cov(Y,X) Var(Y)

What does Σ look like if X and Y are independent?

Σ = (Var(X) 0
(0 Var(Y)

Two-dimensional training data

C1: (10, .1), (5, 0.01), (8, .2), (15, .1), (17, .25), (18, .5), (2, .01), (3, .01), (11, .1), (12, .2)
C2: (2, .1), (0, .2), (0, 1), (1, 1), (10, 2), (0, .01), (2, .1), (4, .5) (1, .5), (0, .75)

Test data: x' = (15, 1)

We use Bayesian decision theory. This means we have to calculate

P(C1|x') and P(C2|x') and select the class with the higher probability.

Assume that P(C1) = .04, P(C2) = .96

P(C1|x') = P(x'|C1)P(C1)

So all we need to do is find P(x'|C1). There are two ways. We could take the tabular approach that we did earlier. Or we can assume that P(x'|C1) is Gaussian. Let us take the Gaussian approach.

If we assume that P(x'|C1) is Gaussian then we can write P(x'|C1) as P(x'|m1, Σ 1). We need to estimate the mean and variance of our Gaussian distribution. Our slides give us the maximum likelihood estimates.

m1 = (10.1, .15)

Σ 1 = (28.5 .6)
(.6 .02)

Inverse of Σ 1 = (.093 -2.72)
(-2.72 129.4)

What is the rule to multiply (a b) with (e f) ?
(c d) (g h)

The result is (ae+bg af+bh)
(ce+dg cf+dh)

Evaluate : (28.5*.093 + .6*-2.72 28.5*-2.72 + .6*129.4) = (1 0)
(.6*.093 + .02*-2.72 .6*-2.72 + .02*129.4) (0 1)

Similarly we have P(x'|C2) = P(x'|m2, Σ 2)

m2 = (2, .62)

Σ 2 = (8.6 1.16)
(1.16 0.33)

Inverse of $\Sigma 2$ = (.22 -.76)
(-.76 5.65)

Now we calculate,

P(x'=(15,1)|m1, $\Sigma 1$) = Gaussian(x'=(15,1), m1, $\Sigma 1$)

and

P(x'=(15,1)|m2, $\Sigma 2$) = Gaussian(x'=(15,1), m2, $\Sigma 2$)
