Lecture 2:

Pattern Recognition and Decision Theory

MSDM 5058

Prepared by S.P. Li

Pattern Recognition and Decision Theory

We discuss two topics in this lecture:

1) How is our study of probability related to the topics of pattern recognition?

2) How can we use pattern recognition in making decision in real applications?

Pattern Recognition

Given a lot of data in the past, how can we make prediction for the future?

Intuitively, we should find some patterns that occur frequently in the past data, and expect these patterns will also occur in the future.

In science, this may lead to the discovery of new physical laws.

In real application, we may be less ambitious, just hope to get some rules in predicting future patterns

Pattern Recognition and Probability

As patterns occur frequently in our database should have higher chance to occur in the future, we should focus on those patterns that occur frequently.

But how does this relate to prediction rule?

This is related to time ordering in the data.

In science, causality is fundamental.

A causes B means that A must occur before B

Thus, causality necessarily implies some time ordering in the data.

On the other hand, a set of data with time ordering may not reveal any causality or at least any causal effect that we can discover easily.

Data preprocessing before Pattern Recognition in time series

Let us now consider real time series data that has some hidden causal relation between some data points.

We like to discover the frequently occurring patterns in our database without questioning the deeper question of causality.

This will be sufficient for our application in predicting the future data, even though we do not have a theory of their causal relation.

We first consider some preprocessing of the data, with some bold assumption that the trend hidden in the data can be filtered out so that we can treat the changes between data points as a stationary process. Of course, *this is a bold assumption*, but this is one of the simplest way to proceed, and later we can improve our analysis.

Simplest Data preprocessing before Pattern Recognition in time series

One of the simplest way to preprocess the time series data $\{s(1), s(2),\ldots, s(N)\}$ before searching for patterns is to use the "daily rate of return" time series $\{x(1), x(2),\ldots, x(N\text{-}1)\}$ with

$$x(t) = \frac{s(t) - s(t-1)}{s(t-1)}$$

We have discussed this in lecture 1. We can also discuss other ways to form the time series $\{x(i), i=1,.., N\text{-}1\}$ from $\{s(1), i=1,.., N\}$.

The objective is to construct a preprocessed* $\{x(i), i = 1,\ldots, N\text{-}1\}$ *so that we can find frequently occurring patterns.

The daily rate of return is the fraction of change of the stock price $s(t)$.

Digitization of Daily Rate of Return time series

If we represent the daily rate of return x with the following rule:

- When $x(i) < -r < 0$, we denote the data at time t_i as D
- When $-r < x(i) < r$, we denote the data at time as H
- When $0 < r \leq x(i)$, we denote the data at time t_i as U

Here we assume $r > 0$ is a small number, e.g. $r = 0.01$ (corresponding to a one percent daily rate of return)

In this way, the real number time series $\{x(1), x(2),\ldots, x(N\text{-}1)\}$ is converted into a symbolic sequence of {U,D,H,…}

For example, for the sequence of daily rate of return {0.002, -0.02, 0.03, 0.002, …}

If we use $r = 0.01$, then this real number sequence can be converted into

{H, D, U, H, …}

==> This is a simple way to digitize the real number sequence.

Association Rule and Data Mining

First let's consider the following prediction rule with two parts:

- *Antecedent or left-hand-side (LHS) X*
- *Consequent or right-hand-side (RHS) Y*

Consider the set of observed data, ***I* = {milk, bread, butter, beer, diapers}**

We can form a table of database containing the items with entry value = 1 means the presence of the item and value = 0 represents the absence of an item

Now, If the observed data is X = (0,1,1,0,0), *it means that the customer bought bread and butter,*

Then, the association rule Y = (1,1,1,0,0) *says that this customer also buys milk.*

Of course, this rule can be wrong, but it makes a prediction that the customer who buys bread and butter may also buy milk.

How to evaluate the usefulness of an association rule?

The Support of an association rule is an indication of how frequently the item set appears in the dataset: the probability of occurrence of X in the observed data set T

$$Support(X) \equiv Pr(X)$$

Confidence is an indication of how often the rule has been found to be true.

The confidence value of a rule, $X \Rightarrow Y$, with respect to a set of transactions T is the proportion of the transactions that contains X which also contains Y

$$\textbf{Confidence of the rule } (X \Rightarrow Y) \equiv \frac{Pr(X \text{ and } Y)}{Pr(X)} = Pr(Y|X)$$

The confidence of an association rule is the conditional probability that given X has been observed, what is the probability that Y will occur.

In the context of prediction rule for time series, this can be stated as

Given the observed past data *X*, what is the probability that *Y* will occur in future.

Datamining association rules from a known database: An example

Consider the following dataset *T* tabulated in a supermarket

Transaction number	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

What is the ***Support*** of X = (*, 1, 1, *, *)? We call * the don't care symbol
Here * means we do not care the value of that particular item
Out of 5 transaction observation, X occurs only once, so that $Pr(X) = 1/5 = 0.2$
For Y = (1, *, *, *, *) What is the Confidence of the rule ($X \Rightarrow Y$) ?
We see from the fourth transaction that

$$\textbf{Confidence of the rule } (X \Rightarrow Y) \equiv \frac{Pr(X \text{ and } Y)}{Pr(X)} = Pr(Y|X) = 1$$

Lift of a rule

$$\textbf{\textit{Lift of the rule}}\ (X \Rightarrow Y) \equiv \frac{\Pr(X \text{ and } Y)}{Pr(X)\Pr(Y)} = \frac{\Pr(Y|X)}{\Pr(Y)}$$

- If the lift of a rule = 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no meaningful rule can be drawn
- If the lift is > 1, it tell us the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.
- If the lift is < 1, it tells us the items are substitute to each other. This means that presence of one item has negative effect on presence of other item and vice versa.

==> Sometimes people also use the name ***Interest of a rule*** for ***Lift of a rule***

Lift of a rule: An example

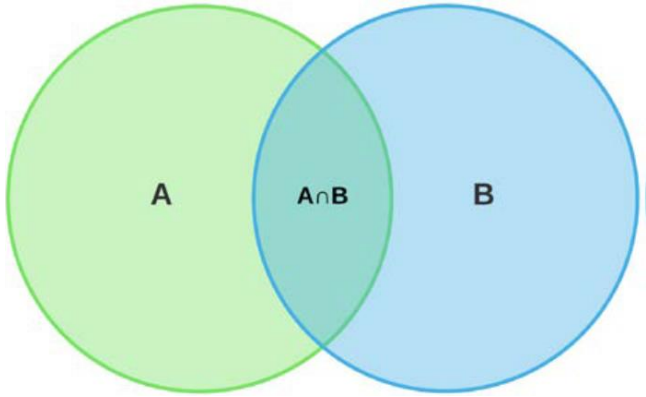
Consider again the following dataset *T* tabulated in a supermarket

Transaction number	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Let's calculate the lift of this rule:
If people buy milk and bread, then they also buy butter.
The rule {milk, bread}: $X = (1, 1, *, *, *) \Rightarrow$ {butter}: $Y = (*, *, 1, *, *)$
Since $P(X) = 2/5$; $P(Y) = 2/5$; $P(X \text{ and } Y) = 1/5$,

$$\textbf{Lift of the rule } (X \Rightarrow Y) \equiv \frac{\Pr(X \text{ and } Y)}{Pr(X)\Pr(Y)} = \frac{\frac{1}{5}}{\left(\frac{2}{5}\right) * \left(\frac{2}{5}\right)} = \frac{5}{4} = 1.25$$

Confidence and RPF of a rule



Confidence is an indication of how often the rule has been found to be true. It is defined by the conditional probability

$$\text{Confidence of the rule } (X \Rightarrow Y) \equiv Pr(Y|X)$$

RPF (Rule Power Factor) measures how intense a rule's items are associated with each other in terms of positive relationship

$$\text{Rule Power Factor of the rule } (X \Rightarrow Y) \equiv Pr(X \text{ and } Y)*Pr(Y|X) = Pr(Y|X)\ P(X)\ Pr(Y|X)$$

$$= \textbf{RPF} \text{ of the rule } (X \Rightarrow Y) = Pr(X \text{ and } Y)* \text{ Confidence of the rule } (X \Rightarrow Y) = P(X)\ \{Pr(Y|X)\}^2$$

- ***RPF*** emphasizes the association between antecedent and consequent of rules.
- ***RPF*** works well even where confidence fails

Note that Probability of (A and B) = $P(A \cap B)$ is the area of the intersection of the Venn diagram of A and the Venn diagram of B

Example to show the importance of RPF measure

Case (1): If item *A* appeared in 20 transactions and *B* in 50 out of total 100 transactions and item *A* and *B* both together appear 15 transactions.

$$\text{Confidence of the rule } (A \Rightarrow B) \equiv Pr(B|A) = Pr(A \text{ and } B)/Pr(A) = 0.15/0.2 = 0.75$$

We get

$$RPF(A \Rightarrow B) = Pr(A \text{ and } B)*Pr(B|A) = 0.15*0.75 = 0.11$$

Case (2): If item *A* appeared in 30 transactions and *B* in 60 out of total 100 transactions and item *A* and *B*, both together appear 20 transactions.

$$\text{Confidence of the rule } (A \Rightarrow B) \equiv Pr(B|A) = Pr(A \text{ and } B)/Pr(A) = 0.2/0.3 = 0.66$$

We get

$$RPF(A \Rightarrow B) = Pr(A \text{ and } B)*Pr(B|A) = 0.66*0.2 = 0.13$$

In Case (2)*, both antecedent and consequent item's occurrences increased individually and together.*

Confidence measure indicates that rule in Case (1) *data* (Conf(1) = 0.75) *is more important than the same rule in Case* (2) *data* (Conf(2) = 0.66)

RPF judges correctly that the rule in Case (2) data is more important: RPF(1) = 0.11 and RPF(2) = 0.13

*The reason is that **RPF** also accounts for the prior probability of occurrence of both the antecedent* (*A*) and *the consequent* (*B*) *of the data.*

Summary on Association rule

We have defined several useful measure of the prediction rule (association rule in general). They are

$$\textbf{\textit{Confidence of the rule}}\ (X \Rightarrow Y) \equiv \frac{Pr(X \text{ and } Y)}{Pr(X)} = Pr(Y|X)$$

$$\textbf{\textit{Lift of the rule}}\ (X \Rightarrow Y) \equiv \frac{\Pr(X \text{ and } Y)}{P\,r(X)\,\Pr(Y)} = \frac{\Pr(Y|X)}{\Pr(Y)}$$

Rule Power Factor of the rule $(X \Rightarrow Y) \equiv Pr(X \text{ and } Y)*Pr(Y|X)$

RPF of the rule $(X \Rightarrow Y)$ = $Pr(X \text{ and } Y)$* Confidence of the rule $(X \Rightarrow Y)$

Prediction rule as Classifier

We consider the relation of prediction rule and classifiers

For time series prediction under the context of association rule, the antecedent *x* is the conditional part, and the consequent *y* is the predicted result.

The simplest consequent part is that the stock price tomorrow will either go up, down, or stay in a small fluctuation range.

In this way, we can also consider the prediction rule as a classifier.

Given the antecedent *x*, what classification one should make (**U**p, **D**own, **H**old) for the next data point? Thus, we can map the problem of finding good association rule as a problem of classification of antecedent and the consequent patterns.

For example, we can consider using past four days data in the stock price time series* (*x*(*t*-3), *x*(*t*-2), *x*(*t*-1), *x*(*t*)) *as the conditional (or antecedent) part and make prediction about tomorrow stock price Y which can be either U, D or H.

Conditional part x (Antecedent)				**Predicted part y, (Consequent)**
X(t-3)	X(t-2)	X(t-1)	X(t)	Y=(U,D,or H)????

Bayes Classifier

Classifiers *are models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. (Here in our example are U,D,H)*

Classifier *needs data to train the classification rules,*
(the same as our need of data for datamining of association rule).
We assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

A fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. Our classifier considers each feature to contribute ***independently*** to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

Bayes rule provides a conditional probability model for classification:

→ Given a set of observed data vector representing some *n* features, we like to classify these data into *K* possible outcomes or classes

Bayes rule in Classifier

For classification, we need to find the posterior probability,

$p(C_k|\vec{x})$ for each of K possible outcomes C_k given that we observed $\vec{x}$.

From Bayes rule, we know that this posterior probability equals to the ratio of (the product of prior and likelihood) over the probability of the occurrence of the observed data

$$p(C_k|\vec{x}) = \frac{p(C_k)p(\vec{x}|C_k)}{p(\vec{x})}, \text{ for each of } K \text{ possible outcomes } C_k$$

with

$p(C_k|\vec{x})$ = posterior probability (It is a conditional probability given the data observed)
$p(C_k)$ = prior probability of the occurrence of class C_k
$p(\vec{x}|C_k)$ = likelihood of $\vec{x}$ given C_k (It is a conditional probability that you will observed the data given that the data is emitted by class C_k)
$p(\vec{x})$ = probability of the occurrence of the observed data $\vec{x}$.

Joint probability in Bayes rule

Note that

$$p(C_k \cap \vec{x}) = p(C_k, \vec{x}) = p(\vec{x}, C_k) = \ p(C_k)p(\vec{x}|C_k) = p(\vec{x})p(C_k|\vec{x})$$

is the joint probability of the occurrence of the observed data $\vec{x}$ and the prior probability of the occurrence of class C_k .

In the case of discrete inputs (indicator or frequency features for discrete events), Bayes classifiers can be considered a way of fitting a probability model that optimizes the joint probability.

Application of Bayes classifier

For example, if we have two classes, (1: Male ; 2: Female) then we can use Bayes classifier to make a prediction:

Given the observed data $\vec{x}$, we predict C_1 (class 1: Male) if the posterior probability

$$p(C_1|\vec{x}) > p(C_2|\vec{x})$$

Sometimes we introduce the log ratio called logit or log-odds defined as

$$\log\left(\frac{p(C_1|\vec{x})}{p(C_2|\vec{x})}\right) = \log\big(p(C_1|\vec{x})\big) - \log\big(p(C_2|\vec{x})\big)$$

When logit (1, 2) > 0, we predict class 1 given observed data $\vec{x}$.

When logit (1, 2) < 0, we predict class 2 in this simple example of two classes.

Constructing a classifier from the probability model is based on the assumption of the independent feature model and Bayes rule to reach a decision rule.

Example of Bayes Classifier

Classify whether a given person is a male or a female based on the measured features. The features include height, weight, and foot size.

Example training set data collected from four males and four females:

Person	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Assumption on Training Set Data

The classifier created from the training set using a Gaussian distribution.
The assumed parameters in the Gaussian (normal) distribution are:

Person	mean (height in ft)	variance (height)	mean (weight lb)	variance (weight)	mean (foot size in Inches)	variance (foot size)
male	5.8550	0.035033	176.25	122.92	11.25	0.9166
female	5.4175	0.097225	132.50	558.33	7.50	1.6667

Let's assume equi-probable classes so that the priors: P(male) = P(female) = 0.5.
This prior probability distribution might be based on our knowledge of frequencies in the larger population, or on frequency in the training set.
Question: Suppose now we have a person with the following observed data.

Height is 6ft, weight=130 pounds, and foot size is 8 inches.

What will the Bayes classifier predict about the gender of this person?

Solution Method by Bayes Classifier

We wish to determine which posterior is greater, male or female. For the classification as male, the posterior is given by

$$p(C_1|\vec{x}) > p(C_2|\vec{x})$$

Here, $\vec{x}$ = (Height, Weight, Footsize) and

$$p(\vec{x}|C_2) = p(Height|C_2)p(Weight|C_2)p(Footsize|C_2)$$

As we assume that the features are independent, the posterior

$$p(C_1|\vec{x}) = p(male|\vec{x}) = \frac{p(\vec{x} \cap C_1)}{p(\vec{x})} = \frac{p(C_1)p(\vec{x}|C_1)}{p(\vec{x})}$$

Often we call $p(\vec{x})$ the evidence, which is a normalization constant and when we compare $p(C_1|\vec{x})$ and $p(C_2|\vec{x})$, we do not need to worry about it.

Solution : Part 1

$$\text{posterior (male)} = \frac{P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male})}{evidence}$$

For the classification as female the posterior is given by

$$\text{posterior (female)} = \frac{P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})}{evidence}$$

The evidence (also termed normalizing constant) may be calculated:

$$\begin{aligned}\text{evidence} = {} & P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male}) \\ & + P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})\end{aligned}$$

$$P(\text{male}) = 0.5$$

$$p(\text{height} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789,$$

where $\mu = 5.855$ and $\sigma^2 = 3.5033 \cdot 10^{-2}$ are the parameters of normal distribution which have been previously determined from the training set. Note that a value greater than 1 is OK here – it is a probability density rather than a probability, because *height* is a continuous variable.

Solution : Part 2

$$p(\text{weight} \mid \text{male}) = 5.9881 \cdot 10^{-6}$$
$$p(\text{foot size} \mid \text{male}) = 1.3112 \cdot 10^{-3}$$
$$\text{posterior numerator (male)} = \text{their product} = 6.1984 \cdot 10^{-9}$$

$$P(\text{female}) = 0.5$$
$$p(\text{height} \mid \text{female}) = 2.2346 \cdot 10^{-1}$$
$$p(\text{weight} \mid \text{female}) = 1.6789 \cdot 10^{-2}$$
$$p(\text{foot size} \mid \text{female}) = 2.8669 \cdot 10^{-1}$$
$$\text{posterior numerator (female)} = \text{their product} = 5.3778 \cdot 10^{-4}$$

Since posterior numerator is greater in the female case, we predict the sample is female.

Decision Theory in Prediction Rule

Given a prediction rule, with given antecedent part (X) and corresponding consequent part (Y), how can we decide on the outcome predicted so as to minimize the cost of our decision?

***Decision theory** is very important as it incorporates the probabilistic nature of the prediction rule with cost of our decision, which ultimately is most important in our decision process.* What we want to know are

1) *What happens if the prediction is correct and we make the right decision to follow the prediction rule?*
2) *What happens if the prediction is correct and we make the wrong decision to and do not follow the prediction rule?*
3) *What happens if the prediction is wrong and we make the right decision to not to follow the prediction rule?*
4) *What happens if the prediction is wrong and we make the wrong decision and follow the prediction rule?*

Example: Decision Theory in the context of communication

How do we find the best communication system so that we minimize the error, maximize the benefit of correct decision?

How do we decide on a signal when there is noise?

How do we know that the answer based on probabilities are the best answer we can provide?

We first discuss signal detection and use of prediction rule by considering a signal (antecedent part X) is observed (may involve noise in the background), and our decision on the consequent part (Y) at cost.

Signal Detection and Parameter Estimation

We first discuss decision theory in the context of communication system: Consider a fixed channel and transmitter,

- How do we design the best receiver?
- What is the optimization criterion?
- What is the structure under this criterion?
- What is the performance of the optimized receiver?

Two parts:

Signal detection + parameter estimation

Optimization Criterion

In communication, the optimization criterion is to minimize the average cost of signal detection or estimator

- *Signal Detection deals with detection of signal in a noisy environment, among other candidate signals*
- *Parameter estimation deals with estimation of some characteristics of the signal*

This is achieved by the ***Bayes Detector and Estimator***

Signal Detection

Problem:

Determine if a particular signal k is present given observed data Z when there is white noise N in the background

Two Hypothesis:

Hypothesis 1: $X = N$ Probability (Hypothesis 1 is true) = p

Hypothesis 2: $X = k + N$ Probability (Hypothesis 2 is true) =1– p

Assume that the noise N is white noise with zero mean and variance σ^2_N

Conditional Probability given Observed Data X

The conditional probability that given Hypothesis 1 is true and the observed data is *x*

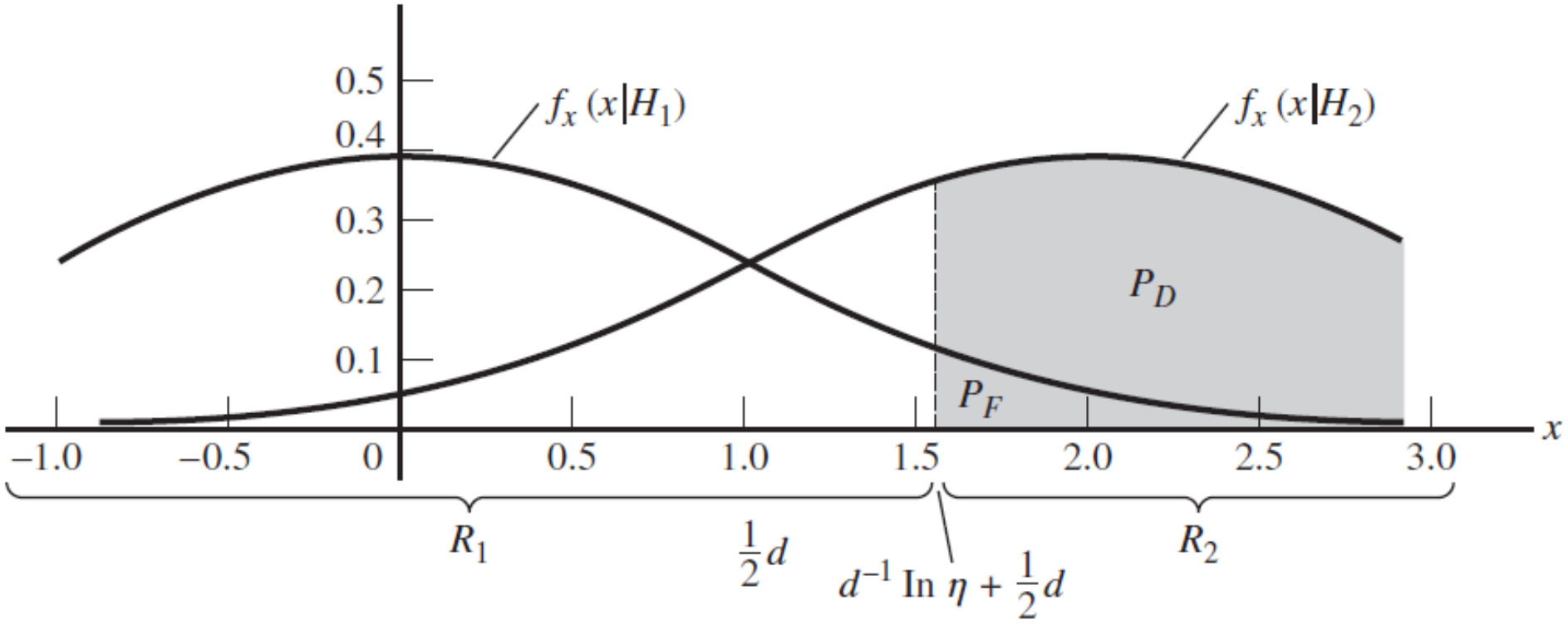
$$f_X(x|H_1) = \frac{e^{-\frac{x^2}{2\sigma_N^2}}}{\sqrt{2\pi\sigma_N^2}}$$

The conditional probability that given Hypothesis 2 is true and the observed data is *x*

$$f_X(x|H_2) = \frac{e^{-\frac{(x-k)^2}{2\sigma_N^2}}}{\sqrt{2\pi\sigma_N^2}}$$

Two-Hypothesis Detection Problem

Partition the observational space *Z* into 2 regions *R1* and *R2* so that if *Z* falls into *R1* we will decide *H1* to be true ($d = k/\sigma_N$)



Decision Problem

We decide *R1* and *R2* so that the average cost of making a decision is minimum!

Note that it may happen that the two regions *R1* and *R2* consists of many segments in the observational space *x*

Cost definition:

c_{11}=cost of deciding in favor of *H1* when *H1* is true
c_{22}=cost of deciding in favor of *H2* when *H2* is true
c_{12}=cost of deciding in favor of *H1* when *H2* is true
c_{21}=cost of deciding in favor of *H2* when *H1* is true

c_{11} and c_{22} are reward for correct decision
c_{12} and c_{21} are penalty for wrong decision

Decision Cost Calculation

What is the cost of decision D given that the hypothesis H1 is true?

Denote this cost as $C(D|H_1)$, then

$$C(D|H_1) = c_{11}P(decide\ H_1|H_1\ is\ true) + c_{21}P(decide\ H_2|H_1\ is\ true)$$

Here,

$$P(decide\ H_1|H_1\ is\ true) = \int_{R_1} f_Z(z|H_1)dz$$

$$P(decide\ H_2|H_1\ is\ true) = \int_{R_2} f_Z(z|H_1)dz$$

Cost of decision when Hypothesis 2 is true

$$C(D|H_2) = c_{12}P(decide\ H_1|H_2\ is\ true) + c_{22}P(decide\ H_2|H_2\ is\ true)$$

Here,

$$P(decide\ H_1|H_2\ is\ true) = \int_{R_1} f_Z(z|H_2)dz$$

$$P(decide\ H_2|H_2\ is\ true) = \int_{R_2} f_Z(z|H_2)dz$$

Average Cost of Decision

The above cost concerns the condition probabilities, so to get the average cost of decision *C*(*D*), we need to compute the average *C*(*D*) with the prior probability ***Pr*** that each of the hypothesis is true:

$$C(D) = Pr(H_1\ is\ true)C(D|H_1) + Pr(H_2\ is\ true)C(D|H_2)$$

Using the following relations:

$$Pr(H_1\ is\ true) + Pr(H_2\ is\ true) = 1$$
$$P(decide\ H_1|H_2\ is\ true) + P(decide\ H_2|H_2\ is\ true) = 1$$

One can simplify *C*(*D*) to get a formula that only involve the integration over the region *R1*.

Analysis I

The final result is $C(D)$ = constant + integral over region *R1*:

$$C(D) = (p_0 c_{21} + q_0 c_{22}) + \int_{R_1} \{I_2 - I_1\} dz$$

Here, $(p_0 c_{21} + q_0 c_{22})$ is a fixed cost and,

$$I_2 = [(c_{12} - c_{22}) q_0 f_Z(z|H_2)] > 0 \text{ since } c_{12} > c_{22}$$
$$I_1 = [(c_{21} - c_{11}) p_0 f_Z(z|H_1)] > 0 \text{ since } c_{21} > c_{11}$$

and $f_Z(z|H) > 0$,

$$Pr(H_1 \; is \; true) = p_0 > 0$$
$$Pr(H_2 \; is \; true) = q_0 = 1 - p_0 > 0$$

First term is the fixed cost and the second term is an integral over a region *R1*.

Analysis II

Since the cost of correct decision is less than the cost of wrong decision, therefore $c_{12} > c_{22}$ and $c_{21} > c_{11}$, and that p_0, q_0 and f_z are probabilities so that they are positive, therefore *I1* and *I2* are both positive.

In order that the cost is minimized, values of the partition of the observed data z into *R1* and *R2* must be made in such a way that

$$C(D) = (p_0 c_{21} + q_0 c_{22}) + \int_{R_1} \{I_2 - I_1\} dz$$

will be reduced if

$\int_{R_1} I_2 dz$ is minimized while $\int_{R_1} I_1 dz$ is maximized

Values of z that give a larger value for I1 than I2 must be given to region R1 and
Values of z that give a smaller value for I1 than I2 must be given to region R2

Simple Decision Rule Based on Bayes Analysis

The proper partition of the region *R1* and *R2* requires careful computation of the integrals, but a simple rule can be made by noting that:

If

$$I_2 = [(c_{12} - c_{22})q_0 f_Z(z|H_2)] < I_1 = [(c_{21} - c_{11})p_0 f_Z(z|H_1)]$$

or,

$$\frac{f_Z(z|H_2)}{f_Z(z|H_1)} < \frac{(c_{21} - c_{11})p_0}{(c_{12} - c_{22})q_0},$$

then choose Hypothesis 1.

If

$$I_2 = [(c_{12} - c_{22})q_0 f_Z(z|H_2)] > I_1 = [(c_{21} - c_{11})p_0 f_Z(z|H_1)]$$

or,

$$\frac{f_Z(z|H_2)}{f_Z(z|H_1)} > \frac{(c_{21} - c_{11})p_0}{(c_{12} - c_{22})q_0},$$

then choose Hypothesis 2.

Likelihood Ratio and Threshold of Bayes Detector

- The likelihood ratio is defined by

$$\Lambda(Z) \cong \frac{f_Z(z|H_2)}{f_Z(z|H_1)}$$

- The threshold is defined by

$$\eta \cong \frac{(c_{21} - c_{11})p_0}{(c_{12} - c_{22})q_0}$$

Performance of Bayes Detector

The Bayes criterion of minimum average cost:

- *Hypothesis 2 is true when*

$$\Lambda(Z) \cong \frac{f_Z(z|H_2)}{f_Z(z|H_1)} > \eta \cong \frac{(c_{21} - c_{11})p_0}{(c_{12} - c_{22})q_0}$$

- *Hypothesis 1 is true when*

$$\Lambda(Z) \cong \frac{f_Z(z|H_2)}{f_Z(z|H_1)} < \eta \cong \frac{(c_{21} - c_{11})p_0}{(c_{12} - c_{22})q_0}$$

Conditional Probability of Wrong Decision

Probability of *False Alarm*: $P_F \cong \int_{R_2} f_Z(z|H_1)dz$

(“***False Alarm***” is also called a ***false positive in medical field***

You receive a positive result for a test (signal of missile attack is above the critical value so you issue an alarm), but actually your conclusion is wrong, there is no missile attack.) ***False alarm*** is also called a ***Type I error*** in statistics.

Probability of *Miss*: $P_M \cong \int_{R_1} f_Z(z|H_2)dz = 1 - \int_{R_2} f_Z(z|H_2)dz$

(“***Miss*** ” is also called a ***false negative in medical field***

You receive a negative result for a test (signal of missile attack is below the critical value so you DO NOT issue an alarm), but actually your conclusion is wrong, there is missile attack.) ***Miss*** is also called a ***Type II error*** in statistics.

False negative is an error in which a test result improperly indicates no presence of a condition (the result is *negative*). An example for a false negative in medicine is a test indicating that a woman is not pregnant whereas she is actually pregnant.

Conditional Probability of Correct Decision

Probability of *Detection* of the missile attack

$$P_D \cong \int_{R_2} f_Z(z|H_2)dz$$

Probability of *correct prediction* of no missile attack

$$P_N = 1 - P_F \cong \int_{R_1} f_Z(z|H_1)dz$$

These terminology refer to radar detection situation,
where Hypothesis *H2* corresponds to the signal-present hypothesis
and Hypothesis *H1* corresponds to no signal, only noise situation.

Interpretation: Examples

H2 = hypothesis: hostile attack by enemy's plane is coming.

H1 = hypothesis that there is only noise far away, no attack.

Probability of ***false alarm*** is that there is no plane attack, but we think that there is, and give out a false alarm

Probability of ***miss*** is that there is actually a plane attack, but we think that there is none, so that we got attacked and we miss the chance of sending out an alarm

Probability of ***detection*** is the chance that we actually detect the attacking plane coming.

Probability of Miss in terms of Likelihood and Threshold

Given that *H2* is true, a wrong decision of *H1* is made

when $\Lambda(Z) < \eta$ (which suggests *H1*).

Thus, the probability that $\Lambda(Z) < \eta$ is satisfied

$$Pr(\Lambda(Z) < \eta) = \int_0^{\eta} f_{\Lambda}(\lambda|H_2)d\lambda$$

is the Probability of Miss (Wrong Detection) $\boldsymbol{P_M}$

False Alarm in terms of Likelihood and Threshold

Given that *H1* is true, a wrong decision of *H2* is made when $\Lambda(Z) > \eta$ (which suggests *H2*).

Thus, the probability that $\Lambda(Z) > \eta$ is satisfied

$$Pr(\Lambda(Z) > \eta) = \int_{\eta}^{\infty} f_{\Lambda}(\lambda|H_1)d\lambda$$

is the Probability of False Alarm $\boldsymbol{P_F}$

Neyman-Pearson Detector

If the costs c_{ij} and the a priori probabilities p_0 are known, then the Bayes detector cost or risk per decision

$$C(D) = p_0 c_{21} + q_0 c_{22} + q_0(c_{12} - c_{22})P_M - p_0(c_{21} - c_{11})(1 - P_F)$$

When the costs and priors are not available, we can then fixed the false alarm probability P_F at some level and maximize the detection probability P_D (or minimize the probability of miss P_M) subject to

$$P_F \leq \alpha$$

This operation defines the ***Neyman-Pearson Detector***

Minimum Probability of Error Detector

If the cost for making right decision is zero,

$$c_{11} = c_{22} = 0$$

and the cost for making wrong decision is 1,

$$c_{21} = c_{12} = 1$$

then the Bayes cost or risk per decision is

$$C(D) = p_0 P_F + q_0 P_M$$

This implies minimizing Bayes cost is equivalent to finding a minimum probability of error for the first term is false alarm and second is miss.

MAP Detector

Maximum a Posteriori (MAP) detector assumes :

$$c_{11} = c_{22} = 0 \text{ and } c_{21} = c_{12}$$

Then the Bayes optimal condition for decision reads

If

$$I_2 = [(c_{12} - c_{22})q_0 f_Z(z|H_2)] < I_1 = [(c_{21} - c_{11})p_0 f_Z(z|H_1)]$$

or, $\frac{f_Z(z|H_2)}{f_Z(z|H_1)} < \frac{p_0}{q_0}$, then choose Hypothesis 1.

If

$$I_2 = [(c_{12} - c_{22})q_0 f_Z(z|H_2)] > I_1 = [(c_{21} - c_{11})p_0 f_Z(z|H_1)]$$

or, $\frac{f_Z(z|H_2)}{f_Z(z|H_1)} > \frac{p_0}{q_0}$, then choose Hypothesis 2.

Simple MAP

If $\frac{f_Z(z|H_2)}{f_Z(z|H_1)} < \frac{p_0}{q_0} \leftrightarrow f_Z(z|H_2)P(H_2) < f_Z(z|H_1)P(H_1) \leftrightarrow \frac{f_Z(z|H_2)P(H_2)}{f_Z(z)} < \frac{f_Z(z|H_1)P(H_1)}{f_Z(z)}$,

then choose Hypothesis 1

Note that we can use Bayes rule to rewrite

$$P(H_1|z)f_Z(z) = f_Z(z|H_1)P(H_1)$$

$$P(H_2|z)f_Z(z) = f_Z(z|H_2)P(H_2)$$

We get the simple MAP condition

If $P(H_2|z) < P(H_1|z)$, then *H1*

If $P(H_2|z) > P(H_1|z)$, then *H2*

Provided that

$$c_{11} = c_{22} = 0 \text{ and } c_{21} = c_{12}$$

Summary on Decision Theory

The most general detector is ***Bayes Optimal detector***.

The ***likelihood and threshold test*** is to simplify the computation of Bayes Risk of Decision or Decision Cost $C(D)$.

The introduction of ***F****alse* ***A****larm*, ***M****iss*, *and* ***D****etection* probability is for easy application of Bayes Optimal Detector.

The ***Neyman-Pearson Detector*** is used for certain level of performance in decision making when none of the cost or priors are known.

The ***Minimum Probability of Error Detector*** can be used if we do not know the cost and only care about minimizing error.

The ***MAP detector*** is used when we do not know the priors and cost, but want to decide based on observations (thus, a posteriori) and minimize risk, so this is equivalent to the minimum probability of error detector and is a simple way of implementing it using only a posteriori probabilities.

Summary of Pattern Recognition

We first discuss

(*1*) *How is our study of probability related to the topics of pattern recognition?*

We introduce prediction rule as a conditional probability involving

Antecedent X (which is observed) and Consequent Y (which is predicted)

$$\boldsymbol{Rule}\ (X \Rightarrow Y) \equiv \frac{Pr(X \text{ and } Y)}{Pr(X)} = Pr(Y|X)$$

We call this ***confidence*** of the rule, and we introduce other measures of the rule such as the lift, the ***RPF*** (rule power factor) to highlight the importance of a particular rule.

This is under the general topics of **ASSOCIATION RULE DATA MINING**.

Summary of Bayes Classifier

- The problem of ASSOCIATION RULE DATA MINING can be mapped into a problem of classification
- In the context of time series prediction, we can first digitize the time series and make the continuous prediction of stock values as a classification of features such as Up, Down, or Hold.
- The problem is casted in the framework of Bayes rule : We like to get $p(C_k|\vec{x})$ which is the conditional probability of the occurrence of class C_k given the occurrence of the observed data $\vec{x}$.
- The Bayes rule relate the posterior probability $p(C_k|\vec{x})$ with the expression

$$p(C_k \cap \vec{x}) = p(C_k, \vec{x}) = p(\vec{x}, C_k) = \ p(C_k)p(\vec{x}|C_k) = p(\vec{x})p(C_k|\vec{x})$$

Summary of the application of this Lecture

The second question is :

How can we use pattern recognition in making decision in real application?

Based on the digitization of the daily rate of return of the stock time series, we map the stock price into a symbolic sequence consisting of (U, D, H) for up, down, hold. The classifier provide a prediction rule in the form of

Conditional part x (Antecedent)				**Predicted part y, (Consequent)**
X(t-3)	X(t-2)	X(t-1)	X(t)	Y=(U,D,or H)????

Here $X(t\text{-}i)$ is the digitized symbols (U, D or H).

After an evaluation of the importance of the rule, we can then compute the cost of the decision on Buy, Sell, and Hold according to Decision Theory.

Quick Review on Normal Random Variables

A random variable is said to be normally distributed with parameters μ and σ^2, and we write $X \sim N(\mu, \sigma^2)$, if its density is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

The mean

$$E[X] = \mu$$

Variance

$$Var(X) = \sigma^2$$

If $X \sim N(\mu, \sigma^2)$, then $a\,X + b \sim N(a\mu+b, a^2\sigma^2)$. One can define a variable

$$Z = \frac{X - \mu}{\sigma}$$

So that $Z \sim N(0, 1)$.

Quick Review on Normal Random Variables

$Z \sim N(0, 1)$ is called the Standard Normal, and has a density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \qquad -\infty < x < \infty$$

with cumulative distribution function

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy, \qquad -\infty < x < \infty$$

Compute the probability for $X < b$,

$$P\{X < b\} = P\left\{\frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right\} = P\left\{Z < \frac{b-\mu}{\sigma}\right\} = \Phi\left(\frac{b-\mu}{\sigma}\right)$$

and,

$$P\{a < X < b\} = P\{X < b\} - P\{X < a\} = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

Quick Review on Normal Random Variables

For any number x,

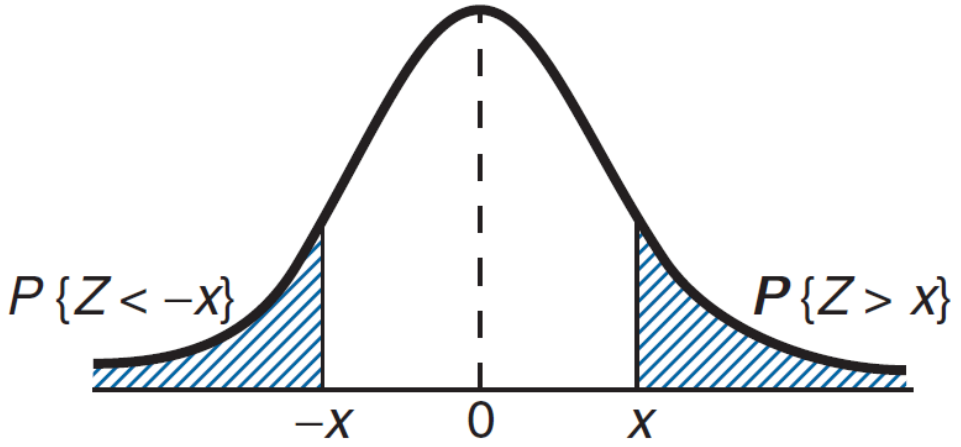
$$P\{Z > x\} = 1 - P\{Z \leq x\} = 1 - \Phi(x)$$

and,

$$P\{Z < -x\} = \Phi(-x)$$

Symmetric Property:

$$\Phi(-x) = 1 - \Phi(x)$$



Quick Review on Normal Random Variables

For $0 < \alpha < 1$, let z_α be a value such that

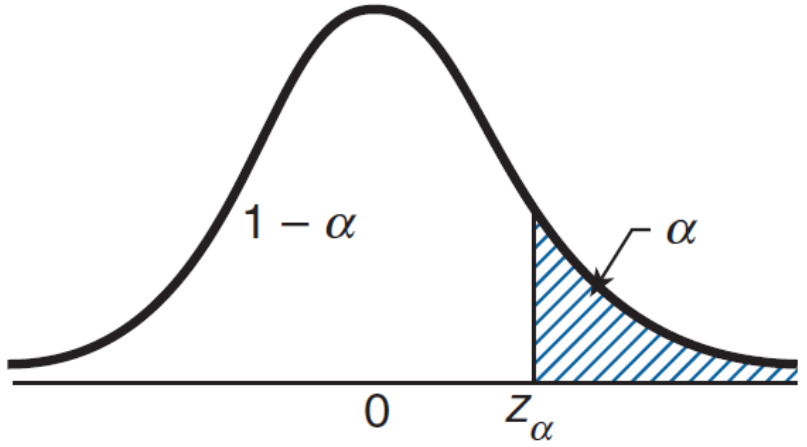
$$P\{Z < z_\alpha\} = 1 - \alpha; \;\; P\{Z > z_\alpha\} = \alpha$$

This means that

$$1 - \Phi(z_\alpha) = \alpha$$

We can also write z_α using α as

$$z_\alpha = \Phi^{-1}(1 - \alpha)$$



Quick Review on Normal Random Variables

An Example:

Suppose that a binary message – either 0 or 1 – must be transmitted by wire from A to B. However, the data sent over the wire are subject to a channel noise disturbance, which can be modelled by a standard normal variable N. To reduce the possibility of error, the value 2 is sent over the wire when the message is 1 and the value -2 is sent when the message is 0. If x, $x = \pm 2$, is the value sent from A, then R, the value received by B, is given by $R = x + N$. When the message is received at location B, the receiver decodes it according to the following rule: If $R \geq 0.5$, then 1 is concluded; If $R < 0.5$, then 0 is concluded.

Two types of errors can occur: (i) message 1 is incorrectly determined to be 0, and (ii) 0 is incorrectly determined to be 1.

The first type of error will occur if the message is 1 and $2 + N < 0.5$, whereas the second will occur if the message is 0 and $-2 + N \geq 0.5$. Hence,

$$P\{\text{error}|\text{message} = 1\} = P\{2 + N < 0.5\} = P\{N < 1.5\} = 1 - \Phi(1.5) \approx 0.0668$$

and,

$$P\{\text{error}|\text{message} = 0\} = P\{-2 + N \geq 0.5\} = P\{N \geq 2.5\} = 1 - \Phi(2.5) \approx 0.0062$$