18CSE751 – Introduction to Machine Learning
Lecture 12-14: Bayesian Decision Theory

Dr.Vani Vasudevan

Professor –CSE, NMIT

UNIT III

Bayesian Learning : Introduction , Classification, Losses and Risks, Discriminant Functions, Utility Theory, Association Rules, Bayes Optimal Classifier, Naïve Bayes Classifier, Bayesian Belief Networks. **Nearest Neighbor Methods**: K-nearest Neighbor Learning, Distance – Weighted Nearest Neighbor Algorithm , Examples (T2-Chapter-3,8 ;T1-Chapters 7, 16, R1- Chapter 6, Chapter 8)

2

BAYESIAN DECISION THEORY

- Discuss probability theory as the framework for making decisions under uncertainty.
- To calculate the probabilities of the classes using bayes rules.
- To discuss how we can make rational decisions among multiple actions to minimize expected risk.
- To discuss learning association rules from data.

3

PROBABILITY AND INFERENCE…

- Result of tossing a coin $\in$ {heads,tails}

Outcome X as a random variable drawn from a probability distribution P(X = x) that specifies the process.

- Random var $X \in \{1,0\}$

X = 1 denotes that the outcome of a toss is heads and X = 0 denotes tails. Such X are bernoulli distributed where the parameter of the distribution p_o is the probability that the outcome is heads:

- Bernoulli: $P\{X=1\} = p_o^{\,X}(1 - p_o)^{(1-X)}$

4

Tossing a coin is a random process because we cannot predict at any toss whether the outcome will be heads or tails. We can only talk about the probability that the outcome of the next toss will be heads or tails. It may be argued that if we have access to extra knowledge such as the exact composition of the coin, its initial position, the force and its direction that is applied to the coin when tossing it, where and how it is caught, and so forth, the exact outcome of the toss can be predicted.

Assume that we are asked to predict the outcome of the next toss. If we know p_o, our prediction will be heads if $p_o > 0.5$ and tails otherwise.
This is because if we choose the more probable case, the probability of error, which is 1 minus the probability of our choice, will be minimum. If this is a fair coin with $p_o = 0.5$, we have no better means of prediction than choosing heads all the time or tossing a fair coin ourselves!

PROBABILITY AND INFERENCE

- Sample: $\mathbf{X} = \{x^t\}_{t=1}^N$
- In the coin tossing example, the sample contains the outcomes of the past N tosses. Then using X, we can estimate p_o,

  Estimation of p_o: $\hat{p}_o = \#\{\text{heads}\}/\#\{\text{tosses}\} = \sum_t x^t / N$
- Numerically using the random variables, x^t is 1 if the outcome of toss t is heads and 0 otherwise
- Given the sample {heads, heads, heads, tails, heads, tails, tails, heads, heads}, we have X = {1, 1, 1, 0, 1, 0, 0, 1, 1} and the estimate is

  $$\hat{p}o = \sum_{t=1} x^t / N = 6/9$$
- Prediction of next toss: Heads if $p_o > ½$, tails otherwise

5

If we do not know P(X) and want to estimate this from a given sample(realm of statistics). We have a sample, X, containing examples drawn from the probability distribution of the observables x^t , denoted as p(x). The aim is to build an approximator to it, ˆp(x), using the sample X

CLASSIFICATION…

Scenario: In a bank, according to their past transactions, some **customers** are **low-risk** in that they paid back their loans and the bank profited from them and other customers are **high-risk** in that they defaulted. Analysing this data, we would like to learn the class "high-risk customer" so that in the future, when there is a new application for a loan, we can check whether that person obeys the class description or not and thus accept or reject the application. Using our knowledge of the application, let us say that we decide that there are **two pieces of information that are observable**. We observe them because we have reason to believe that they give us an idea about the credibility of a customer. Let us say, for example, we **observe customer's yearly income and savings,** which we represent by **two random variables X_1 and X_2.**

CLASSIFICATION

- Credit scoring: inputs are income and savings.
  Output is low-risk vs high-risk
- Input: vector of observables : $\mathbf{x} = [x_1, x_2]^T$,output: $C = \{0,1\}$
- Prediction:

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C=1 \mid x_1, x_2) > 0.5 \\ C = 0 \text{ otherwise} \end{cases}$$

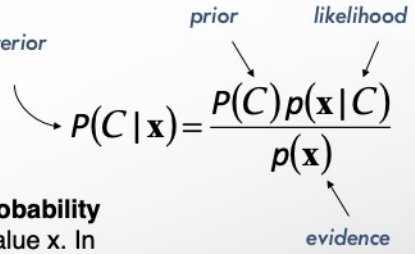
or

$$\text{choose} \begin{cases} C = 1 \text{ if } P(C=1 \mid x_1, x_2) > P(C=0 \mid x_1, x_2) \\ C = 0 \text{ otherwise} \end{cases}$$

7

The probability of error is 1 − max(P(C=1|x1,x2),P(c=0|x1,x2))

BAYES' RULE...



P(C = 1) is called the **prior probability** that C takes the value 1, In the example, it corresponds to the probability that a customer is high risk.

p(xIC) is called the **class likelihood** and is the **conditional probability** that an event belonging to C has the associated observation value x. In our case, p(x1, x2IC = 1) is the probability that a high-risk customer has his or her X1 = x1 and X2 = x2.

p(x), **the evidence**, is the marginal probability that an observation x is seen, regardless of whether it is a positive or negative example

8

BAYES' RULE...

$$P(C=0)+P(C=1)=1$$
$$p(\mathbf{x})=p(\mathbf{x}\mid C=1)P(C=1)+p(\mathbf{x}\mid C=0)P(C=0)$$
$$p(C=0\mid\mathbf{x})+P(C=1\mid\mathbf{x})=1$$

9

BAYES' RULE: K>2 CLASSES

In the general case, we have K mutually exclusive an exhaustive classes; Ci, i = 1, . . . , K;
For example, In optical digit recognition, the input is a bitmap image and there are ten classes.
We have the prior probabilities satisfying

$$P(C_i) \geq 0 \text{ and } \sum_{i=1}^{K} P(C_i) = 1$$

choose C_i if $P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$

$$P(C_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid C_i) P(C_i)}{p(\mathbf{x})}$$

The posterior probability of class Ci can be calculated as

$$= \frac{p(\mathbf{x} \mid C_i) P(C_i)}{\sum_{k=1}^{K} p(\mathbf{x} \mid C_k) P(C_k)}$$

and for minimum error, the Bayes' classifier chooses the class with the highest posterior probability;

Choose C_i : If $(P(C_i|X) = max_K P(C_k|X)$

10

LOSSES AND RISKS…

- It may be the case that decisions are not equally good or costly. A financial institution when making a decision for a loan applicant should take into account the potential gain and loss as well.
  - **An accepted low-risk applicant increases profit, while a rejected high-risk applicant decreases loss.**
  - **The loss for a high-risk applicant erroneously accepted may be different from the potential gain for an erroneously rejected low-risk applicant.**
- The situation is much more critical and far from symmetry in other domains like medical diagnosis or earthquake prediction.

11

LOSSES AND RISKS…

- ACTIONS: α_i
- LOSS OF α_i WHEN THE STATE İS C_K : λ_{iK}
- EXPECTED RISK (DUDA AND HART, 1973)

$$R(\alpha_i \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \mathbf{x})$$ AND WE CHOOSE THE ACTION WITH MINIMUM RISK

$$\text{choose } \alpha_i \text{ if } R(\alpha_i \mid \mathbf{x}) = \min_k R(\alpha_k \mid \mathbf{x})$$

LOSSES AND RISKS: 0/1 LOSS…

Let us define K actions α_i, i = 1, . . . , K, where α_i is the action of assigning x to C_i.

In the special case of the 0/1 loss where $\lambda_{ik} = \begin{cases} 0 \text{ if } i = k \\ 1 \text{ if } i \neq k \end{cases}$

All correct decisions have no loss and all errors are equally costly. The risk of taking action αi is

$$R(\alpha_i \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda_{ik} P(C_k \mid \mathbf{x}) = \sum_{k \neq i} P(C_k \mid \mathbf{x}) = 1 - P(C_i \mid \mathbf{x})$$

For minimum risk, choose the most probable class

13

LOSSES AND RISKS: REJECT…

- In some applications, wrong decisions—namely, misclassifications— may have very high cost, and it is generally required that a more complex— for example, manual—decision is made if the automatic system has low certainty of its decision.
- For example,
  - if we are using an optical digit recognizer to read postal codes on envelopes, wrongly recognizing the code causes the envelope to be sent to a wrong destination.
- In such reject a case, we define an additional action of reject or doubt, α_{K+1}, with α_i, i = 1, . . . , K, being the usual actions of deciding on classes Ci, i = 1, . . . , K

14

LOSSES AND RISKS: REJECT

A possible loss function is

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ \lambda & \text{if } i = K+1, \quad 0 < \lambda < 1 \\ 1 & \text{otherwise} \end{cases}$$

is the loss incurred for choosing the (K + 1)st action of reject. Then the risk of reject is

$$R(\alpha_{K+1} \mid \mathbf{x}) = \sum_{k=1}^{K} \lambda P(C_k \mid \mathbf{x}) = \lambda$$

and the risk of choosing class C_i is

$$R(\alpha_i \mid \mathbf{x}) = \sum_{k \neq i} P(C_k \mid \mathbf{x}) = 1 - P(C_i \mid \mathbf{x})$$

The optimal decision rule is to

choose C_i if $P(C_i \mid \mathbf{x}) > P(C_k \mid \mathbf{x}) \ \forall k \neq i$ and $P(C_i \mid \mathbf{x}) > 1 - \lambda$

reject otherwise

15

DISCRIMINANT FUNCTIONS…

Classification can also be seen as implementing a set of *discriminant functions*, $g_i(\boldsymbol{x}), i = 1, \ldots, K$, such that we

choose C_i if $g_i(\boldsymbol{x}) = \max_k g_k(\boldsymbol{x})$

We can represent the Bayes' classifier in this way by setting

$g_i(\boldsymbol{x}) = -R(\alpha_i|\boldsymbol{x})$

and the maximum discriminant function corresponds to minimum conditional risk. When we use the 0/1 loss function, we have

$g_i(\boldsymbol{x}) = P(C_i|\boldsymbol{x})$

or ignoring the common normalizing term, $p(\boldsymbol{x})$, we can write

$g_i(\boldsymbol{x}) = p(\boldsymbol{x}|C_i)P(C_i)$

16

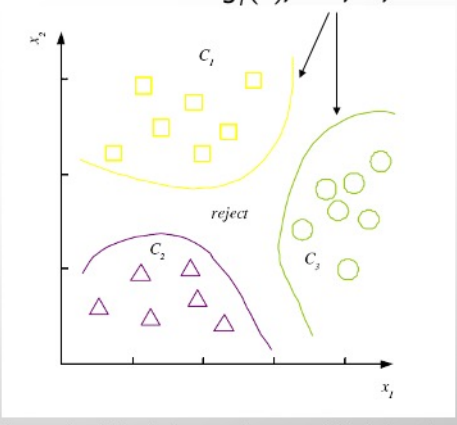
DISCRIMINANT FUNCTIONS…

This divides the feature space into K decision regions $\mathcal{R}_1,\ldots,\mathcal{R}_K$

$$\mathcal{R}_i = \{\mathbf{x} \mid g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$

The regions are separated by decision boundaries, surfaces in feature space where ties occur among the largest discriminant functions

$g_i(\mathbf{x}), i = 1,\ldots,K$



Example of decision regions and decision boundaries

17

$K=2$ CLASSES

- DICHOTOMIZER ($K=2$) VS POLYCHOTOMIZER ($K>2$)
- When there are two classes, we can define a single discriminant
- $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$ $\quad \text{choose} \begin{cases} C_1 \text{ if } g(\mathbf{x}) > 0 \\ C_2 \text{ otherwise} \end{cases}$
- An example is a two-class learning problem where the positive examples can be taken as C_1 and the negative examples as C_2. When $K = 2$,
- the classification system is a **dichotomizer** and for $K \geq 3$, it is a **polychotomizer**

18

UTILITY THEORY

- PROB OF STATE K GIVEN $\mathbf{X}$: $P(S_K \mid \mathbf{X})$
- UTILITY OF α_i WHEN STATE IS K: U_{ik}
- EXPECTED UTILITY:

$$EU(\alpha_i \mid \mathbf{x}) = \sum_k U_{ik} P(S_k \mid \mathbf{x})$$

$$\text{Choose } \alpha_i \text{ if } EU(\alpha_i \mid \mathbf{x}) = \max_j EU(\alpha_j \mid \mathbf{x})$$

19

BAYES CLASSIFIER

- A probabilistic framework for solving classification problems
- Conditional probability:

$$P(C \mid A) = \frac{P(A,C)}{P(A)}$$

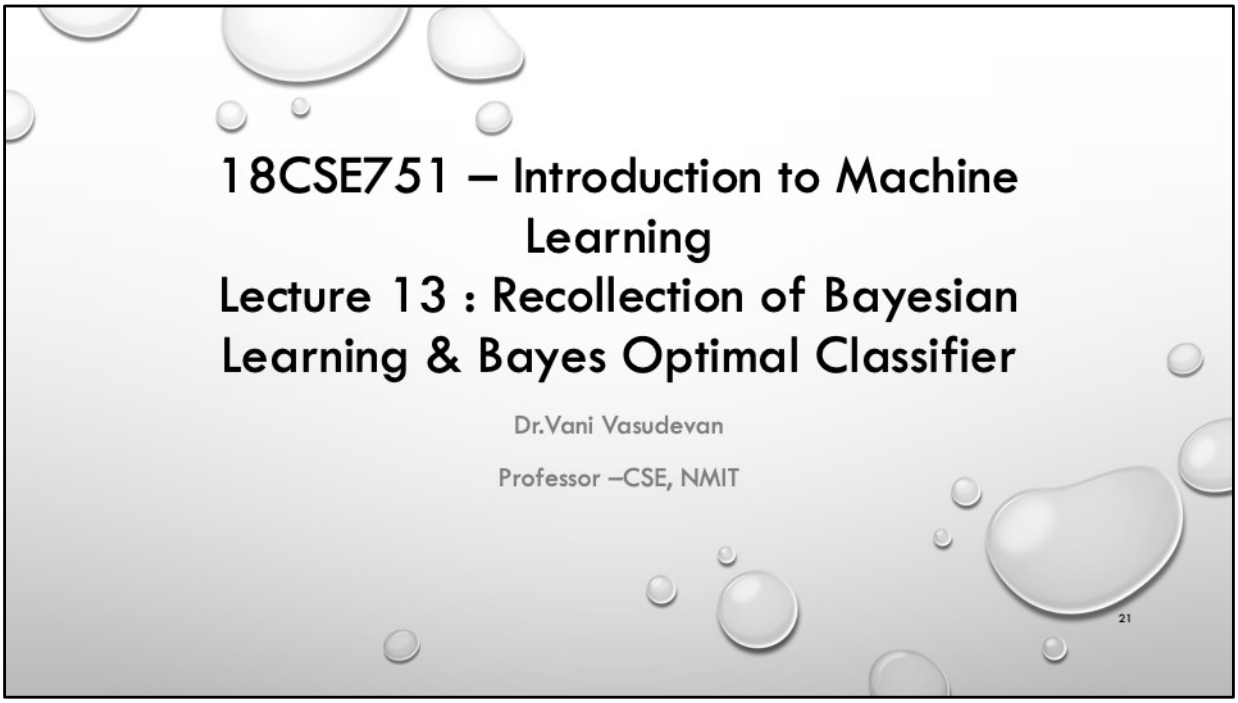
$$P(A \mid C) = \frac{P(A,C)}{P(C)}$$

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**

- Bayes theorem:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

20



Source : Tom m. Mitchell “Machine Learning”, Mc graw Hill publication

BAYESIAN LEARNING…

- Bayesian learning methods(BLMs) are relevant to our study of ML for two different reasons:

1. BLMs that **calculate explicit probabilities for hypothesis,** such as naïve bayes classifier are among the **most practical approaches to certain type of learning problems**. (Problem: classify text documents such as e-news articles)
2. **Provide useful perspective for understanding many learning algorithms that do not explicitly manipulate probabilities.** Bayesian perspective used to analyze the inductive bias of decision tree learning algorithms that favor short decision trees

22

Lecture 13

BAYESIAN LEARNING…

Features of Bayes Learning Methods include:

1. Each observed training example can incrementally decrease or increase the estimated probability that a hypothesis is correct. This provides a more flexible approach to learning than algorithms that eliminate a hypothesis if it is found to be inconsistent with any single example.

23

BAYESIAN LEARNING…

Features of Bayes Learning Methods include:

2. Prior knowledge can be combined with observed data to determine the final probability of a hypothesis. In Bayesian Learning, prior knowledge is provided by asserting

(1) **a prior probability for each candidate hypothesis, and**

(2) **a probability distribution over observed data for each possible hypothesis.**

24

BAYESIAN LEARNING…

Features of Bayes Learning Methods include:

3. Bayesian methods can accommodate hypotheses that make probabilistic predictions (e.g., hypotheses such as "this pneumonia patient has a 93% chance of complete recovery").

4. New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.

25

BAYESIAN LEARNING…

Features of Bayes Learning Methods include:

5. Even in cases where bayesian methods prove computationally intractable, they **can provide a standard of optimal decision making against which other practical methods can be measured.**

26

BAYESIAN LEARNING…

- Difficulties in applying bayesian methods

1. It **require initial knowledge of many probabilities**. When these **probabilities** are not known in advance they are often **estimated** based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.
2. The **significant computational cost required to determine the bayes optimal hypothesis** in the general case (linear in the number of candidate hypotheses). In certain specialized situations, this computational cost can be significantly reduced.

27

MAP HYPOTHESIS…

- **Maximum A posteriori(MAP)** : In many learning scenarios, the **learner considers some set of candidate hypotheses H** and is interested in **finding the most probable hypothesis h ∈ H**
- Given the observed data D (or at least one of the maximally probable if there are several). Any such maximally probable hypothesis is called a **Maximum A Posteriori (MAP) hypothesis**.
- We can determine the MAP hypotheses h_{MAP} by using Bayes theorem to calculate the posterior probability of each candidate hypothesis.

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h|D)$$
$$= \underset{h \in H}{\operatorname{argmax}} \frac{P(D|h)P(h)}{P(D)}$$
$$= \underset{h \in H}{\operatorname{argmax}} P(D|h)P(h)$$

Note: In final step, the term P(D) was dropped because it is a constant independent of h.

Discriminant analysis

MAP HYPOTHESIS

- In some cases, assume that every **hypothesis in H is equally probable a priori** ($P(hi) = P(h_j)\ for\ all\ hi\ and\ h_j\ in\ H$).
- The equation is simplified to the term $P(D|h)$ to find the most probable hypothesis.
- **$P(D|h)$ is often called the likelihood of the data D given h, and any hypothesis that maximizes $P(D|h)$ is called a maximum likelihood (ML) hypothesis,** h_{ML}.

$$h_{ML} \equiv \underset{h \in H}{\operatorname{argmax}}\ P(D|h)$$

Where data D as training examples of some target function and referring to H as the space of candidate target functions.

29

AN EXAMPLE…

- To illustrate bayes rule, consider a medical diagnosis problem in which there are two alternative hypotheses:

(1) that the patients has a particular form of cancer.

(2) that the patient does not.

30

AN EXAMPLE…

With two possible outcomes: $\oplus$ (positive) and $\ominus$ (negative).

Prior knowledge over the entire population of people only .008 have this disease. Furthermore, the lab test is only an imperfect indicator of the disease.

The test returns a **correct positive result in only 98%** of the cases in which the disease is present and **a correct negative result in only 97%** of the cases in which the disease is not present. In other cases, the test returns the opposite result.

31

AN EXAMPLE…

It can be summarized by the following probabilities:

$$P(cancer) = .008, \quad P(\neg cancer) = .992$$
$$P(\oplus|cancer) = .98, \quad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03, \quad P(\ominus|\neg cancer) = .97$$

32

AN EXAMPLE…

- Suppose observe a new patient for whom the **lab test returns a positive result**. Should we diagnose the patient as having cancer or not?
- The MAP hypothesis can be found using equation

$$h_{MAP} \equiv \underset{h \in H}{\text{argmax}}\, P(h|D)$$
$$= \underset{h \in H}{\text{argmax}}\, \frac{P(D|h)P(h)}{P(D)}$$
$$= \underset{h \in H}{\text{argmax}}\, P(D|h)P(h)$$

$$P(\oplus|cancer)P(cancer) = (.98).008 = .0078$$
$$P(\oplus|\neg cancer)P(\neg cancer) = (.03).992 = .0298$$

33

Thus,$h_{Map} = \neg cancer$. The exact posterior probabilities can also be determined by normalizing the above quantities so that they sum to 1 (e.g., $0.0078/(0.0078 * .0298)$ = 0.21. This step is warranted because Bayes theorem states that the posterior probabilities are just the above quantities divided by the probability of the data, P(Positives). Although P(Positives) was not provided directly as part of the problem statement, we can calculate it in this fashion because we know that P(cancer| Positives) and P($\neg cancer$|Positives) must sum to 1 (i.e., either the patient has cancer or they do not).

Notice that while the posterior probability of cancer is significantly higher than its prior probability, the most probable hypothesis is still that the patient does not have cancer.

SUMMARY OF BASIC PROBABILITY FORMULAS

- *Product rule*: probability $P(A \wedge B)$ of a conjunction of two events A and B

$$P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$$

- *Sum rule*: probability of a disjunction of two events A and B

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- *Bayes theorem*: the posterior probability $P(h|D)$ of h given D

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- *Theorem of total probability*: if events $A_1, \ldots, A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$, then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

34

BAYES OPTIMAL CLASSIFIER…

- So far we have considered the question "what is the **most probable hypothesis** given the training data?
- The most significant and closely related question is : **"what is the most probable classification of the new instance given the training data?**
- Although it may seem that this second question can be answered by simply applying the MAP hypothesis to the new instance, in fact it is possible to do better.
- To develop some intuitions, consider a hypothesis space containing three **hypotheses, h_1, h_2, and h_3. Suppose that the posterior probabilities of these hypotheses given the training data are .4, .3, and .3 respectively. Thus, h_1 is the MAP hypothesis.**

35

BAYES OPTIMAL CLASSIFIER…

- **Suppose a new instance x is encountered, which is classified positive by h1, but negative by h2 and h3.**
- Taking all hypotheses into account, **the probability that x is positive is .4** (the probability associated with h_1), **the probability that it is negative is therefore .6.**
- **The most probable classification (negative) in this case is different from the classification generated by the map hypothesis.**

36

BAYES OPTIMAL CLASSIFIER…

- **The most probable classification of the new instance is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities.**
- If the possible classification of the new example can take on any value v_j from some set V, then the probability $P(v_j|D)$ that the correct classification for the new instance is v_j is

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

37

Bayes optimal classification:

$$\underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) \tag{6.18}$$

To illustrate in terms of the above example, the set of possible classifications of the new instance is $V = \{\oplus, \ominus\}$, and

$$P(h_1|D) = .4,\ P(\ominus|h_1) = 0,\ P(\oplus|h_1) = 1$$
$$P(h_2|D) = .3,\ P(\ominus|h_2) = 1,\ P(\oplus|h_2) = 0$$
$$P(h_3|D) = .3,\ P(\ominus|h_3) = 1,\ P(\oplus|h_3) = 0$$

therefore

$$\sum_{h_i \in H} P(\oplus|h_i)P(h_i|D) = .4$$
$$\sum_{h_i \in H} P(\ominus|h_i)P(h_i|D) = .6$$

and

$$\underset{v_j \in \{\oplus, \ominus\}}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = \ominus$$

BAYES OPTIMAL CLASSIFIER…

- The optimal classification of the new instance is the value v_j, for which $P(vj|D)$ is maximum.

38

BAYES OPTIMAL CLASSIFIER

- Any system that classifies new instances according to equation (6.18) is called a bayes optimal classifier, or bayes optimal learner.
- **No other classification method using the same hypothesis space and same prior knowledge can outperform this method on average.**
- This method maximizes the probability that the new instance is classified correctly, given the available data, hypothesis space, and prior probabilities over the hypotheses.

39

BAYESIAN CLASSIFIERS

- Consider each attribute and class label as random variables
- Given a record with attributes (A_1, A_2,…,A_N)
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes P(C | A_1, A_2,…,A_N)
- Can we estimate P(C| A_1, A_2,…,A_N) directly from data?

40

BAYESIAN CLASSIFIERS

- Approach:
  - Compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the bayes theorem

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - Choose value of C that maximizes
    $P(C \mid A_1, A_2, \ldots, A_n)$
  - Equivalent to choosing value of C that maximizes
    $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$

41

- How to estimate $P(A_1, A_2, \ldots, A_N \mid C)$?

NAÏVE BAYES CLASSIFIER

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_N \mid C) = P(A_1 \mid C_J)\, P(A_2 \mid C_J) \ldots P(A_N \mid C_J)$
  - Can estimate $P(A_I \mid C_J)$ for all A_I AND C_J.
  - New point is classified to C_J IF $\mathbf{P(C_J)} \prod \mathbf{P(A_I \mid C_J)}$ is maximal.

42

HOW TO ESTIMATE PROBABILITIES FROM DATA?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- CLASS: $P(C) = N_C/N$
  - E.G., $P(NO) = 7/10$, $P(YES) = 3/10$
- FOR DISCRETE ATTRIBUTES:

  $P(A_I \mid C_K) = |A_{IK}| / N_C$

  - WHERE $|A_{IK}|$ IS NUMBER OF ^{k} INSTANCES HAVING ATTRIBUTE A_I AND BELONGS TO CLASS C_K
  - EXAMPLES:

    $P(STATUS=MARRIED \mid NO) = 4/7$
    $P(REFUND=YES \mid YES)=0$

43

HOW TO ESTIMATE PROBABILITIES FROM DATA?

- FOR CONTINUOUS ATTRIBUTES:
  - DISCRETIZE THE RANGE INTO BINS
    - ONE ORDINAL ATTRIBUTE PER BIN
    - VIOLATES INDEPENDENCE ASSUMPTION
  - TWO-WAY SPLIT: (A < V) OR (A > V)
    - CHOOSE ONLY ONE OF THE TWO SPLITS AS NEW ATTRIBUTE
  - PROBABILITY DENSITY ESTIMATION:
    - ASSUME ATTRIBUTE FOLLOWS A NORMAL DISTRIBUTION
    - USE DATA TO ESTIMATE PARAMETERS OF DISTRIBUTION (E.G., MEAN AND STANDARD DEVIATION)
    - ONCE PROBABILITY DISTRIBUTION IS KNOWN, CAN USE IT TO ESTIMATE THE CONDITIONAL PROBABILITY $P(A_i \mid C)$

HOW TO ESTIMATE PROBABILITIES FROM DATA?

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

- NORMAL DISTRIBUTION:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - ONE FOR EACH (A_i,C_i) PAIR

- FOR (INCOME, CLASS=NO):
  - IF CLASS=NO
    - SAMPLE MEAN = 110
    - SAMPLE VARIANCE = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

45

EXAMPLE OF NAÏVE BAYES CLASSIFIER

Given a Test Record:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No)
  × P(Married| Class=No)
  × P(Income=120K| Class=No)
  = 4/7 × 4/7 × 0.0072 = 0.0024

- P(X|Class=Yes) = P(Refund=No| Class=Yes)
  × P(Married| Class=Yes)
  × P(Income=120K| Class=Yes)
  = 1 × 0 × 1.2 × 10^{-9} = 0

Since P(X|No)P(No) > P(X|Yes)P(Yes)

Therefore P(No|X) > P(Yes|X)

=> Class = No

46

NAÏVE BAYES CLASSIFIER

- IF ONE OF THE CONDITIONAL PROBABILITY IS ZERO, THEN THE ENTIRE EXPRESSION BECOMES ZERO
- PROBABILITY ESTIMATION:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m-estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: number of classes

p: prior probability

m: parameter

47

EXAMPLE OF NAÏVE BAYES CLASSIFIER

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

48

NAÏVE BAYES (SUMMARY)

- Robust to isolated noise points
- Handle missing values by ignoring the instance during probability estimate calculations
- Robust to irrelevant attributes
- Independence assumption may not hold for some attributes
  - Use other techniques such as bayesian belief networks (BBN)

49

BAYESIAN NETWORK…

- PROBABILISTIC GRAPHICAL MODELS
  - USES GRAPH THEORY
  - NODE – REPRESENTS RANDOM VARIABLE (NODES – DISCRETE , IF CONTINUOUS THEN, DISCRETIZE)
  - LINK – CONNECTION BETWEEN RANDOM VARIABLE. IF THERE IS NO LINK THEN, VARIABLES ARE INDEPENDENT
  - TWO NODES COULD BE LINKED THROUGH A THIRD NODE.

50

BAYESIAN NETWORK…

C is conditionally independent of B, given A

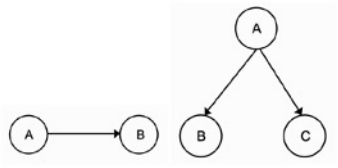


FIGURE 16.1 Two simple graphical models. The arrows denote causal relationships between nodes that represent features.

1. Use directed links: because these relationships are not symmetrical (unless the variables are independent, in which case there is no link).

The one on the left of figure 16.1, mean 'A' causes 'B' (there may be several variables that are all involved in causing B)

2. The probability of A and B is the same as the probability of A times the probability of B conditioned on A: $p(a, b) = p(b|a)p(a)$.

If there is no direct link between two nodes, then they are conditionally independent of each other.

51

BAYESIAN NETWORK…

3. The conditional probability table for each variable. This specifies what the probabilities are for each of the nodes, conditioned on any nodes that are its parents.

- For calculating p(a, b), a distribution table for P(a) and one for p(b|a) are needed.
- The basic concept of the graphical model is very simple
- It makes it more amazing by producing a powerful set of tools for understanding and creating Machine Learning algorithms.
  - Most general model is the **Bayesian Belief Network or Bayesian Network**

BAYESIAN NETWORK …

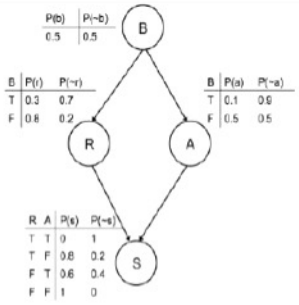


FIGURE 16.2 The sample graphical model. 'B' denotes a node stating whether the exam was boring, 'R' whether or not you revised, 'A' whether or not you attended lectures, and 'S' whether or not you will be scared before the exam.

- DIRECTED GRAPHS WITH NO CYCLES – DAGS (DIRECTED ACYCLIC GRAPHS)
- FOR GRAPHICAL MODELS, WHEN THEY ARE PAIRED WITH THE CONDITIONAL PROBABILITY TABLES, THEY ARE CALLED BAYESIAN NETWORKS.

53

BAYESIAN NETWORK …

- Example: exam fear

S : determine whether you will be scared before an exam based on

(1) whether the course was boring ('B'),

(2) which was the key factor you used to decide whether to attend lectures ('A') and revise ('R').

- **Use the graph to perform inference to decide the likelihood of you being scared before the exam ('S').**

54

BAYESIAN NETWORK …

Two kinds of inferences:

1. Top-down, 2. Bottom-up

- If a set of observations that can be used to predict an unknown outcome, then it is **top-down inference or prediction**
  - If the outcome is known, but the causes are hidden, then it is **bottom-up inference or diagnosis.**
  - Either way, **calculate the values of the hidden (unknown) nodes given information about the observed nodes.**

55

BAYESIAN NETWORK …

- FOR THE GIVEN PROBLEM : PREDICTING WHETHER YOU WILL BE SCARED BEFORE THE EXAM, => **IT IS THE OUTCOME THAT IS HIDDEN.**
- , THE COMPUTATION FOR **EXAM FEAR:**

$$
\begin{aligned}
P(s) &= \sum_{b,r,a} P(b,r,a,s) \\
&= \sum_{b,r,a} P(b) \times P(r|b) \times P(a|b) \times P(s|r,a) \\
&= \sum_{b} P(b) \times \sum_{r,a} P(r|b) \times P(a|b) \times P(s|r,a).
\end{aligned}
$$

56

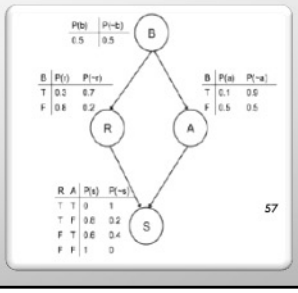
BAYESIAN NETWORK

- If attended lectures and revised are known, then NO need to know whether the course was boring
- ignore the p(b) terms if you know the course is boring, and just need to sum up the probabilities for r and a.

P(S)

$$= \sum P(b) \times \sum P(r|b) \times P(a|b) \times P(s|r,a).$$

- P(S) = **0.3 X0.1 X0** + **0.3 X0.9 X0.8** + **0.7 X0.1 X0.6** + **0.7 X0.9 X1** = 0.328.



57

ASSOCIATION RULES

- ASSOCIATION RULE: $X \rightarrow Y$
- *PEOPLE WHO BUY/CLICK/VISIT/ENJOY X ARE ALSO LIKELY TO BUY/CLICK/VISIT/ENJOY Y.*
- A RULE IMPLIES ASSOCIATION, NOT NECESSARILY CAUSATION.

58

ASSOCIATION MEASURES

- SUPPORT ($X \rightarrow Y$): $$P(X,Y) = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers}\}}$$
- CONFIDENCE ($X \rightarrow Y$): $$P(Y \mid X) = \frac{P(X,Y)}{P(X)} = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$
- LIFT ($X \rightarrow Y$): $$= \frac{P(X,Y)}{P(X)P(Y)} = \frac{P(Y \mid X)}{P(Y)}$$

59

APRIORI ALGORİTHM (AGRAWAL ET AL., 1996)

- FOR (X,Y,Z), A 3-İTEM SET, TO BE FREQUENT (HAVE ENOUGH SUPPORT), (X,Y), (X,Z), AND (Y,Z) SHOULD BE FREQUENT.
- IF (X,Y) İS NOT FREQUENT, NONE OF İTS SUPERSETS CAN BE FREQUENT.
- ONCE WE FİND THE FREQUENT K-İTEM SETS, WE CONVERT THEM TO RULES: X, Y $\rightarrow$ Z, ...

  AND X $\rightarrow$ Y, Z, ...

60

K- NEAREST NEIGHBORS CLASSIFICATION

- K NEAREST NEIGHBORS IS A SIMPLE ALGORITHM THAT STORES ALL AVAILABLE CASES AND CLASSIFIES NEW CASES BASED ON A SIMILARITY MEASURE (E.G., DISTANCE FUNCTIONS). KNN HAS BEEN USED IN STATISTICAL ESTIMATION AND PATTERN RECOGNITION ALREADY IN THE BEGINNING OF 1970'S AS A NON-PARAMETRIC TECHNIQUE.

61

KNN ALGORITHM

- A CASE IS CLASSIFIED BY A MAJORITY VOTE OF ITS NEIGHBORS, WITH THE CASE BEING ASSIGNED TO THE CLASS MOST COMMON AMONGST ITS K NEAREST NEIGHBORS MEASURED BY A DISTANCE FUNCTION. IF K = 1, THEN THE CASE IS SIMPLY ASSIGNED TO THE CLASS OF ITS NEAREST NEIGHBOR.

Distance functions

Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

KNN ALGORITHM

- IT SHOULD ALSO BE NOTED THAT ALL THREE DISTANCE MEASURES ARE ONLY VALID FOR CONTINUOUS VARIABLES.
- IN THE INSTANCE OF CATEGORICAL VARIABLES THE HAMMING DISTANCE MUST BE USED.
- IT ALSO BRINGS UP THE ISSUE OF STANDARDIZATION OF THE NUMERICAL VARIABLES BETWEEN 0 AND 1 WHEN THERE IS A MIXTURE OF NUMERICAL AND CATEGORICAL VARIABLES IN THE DATASET.

KNN ALGORITHM

- CHOOSING THE OPTIMAL VALUE FOR K IS BEST DONE BY FIRST INSPECTING THE DATA. IN GENERAL, A LARGE K VALUE IS MORE PRECISE AS IT REDUCES THE OVERALL NOISE BUT THERE IS NO GUARANTEE. CROSS-VALIDATION IS ANOTHER WAY TO RETROSPECTIVELY DETERMINE A GOOD K VALUE BY USING AN INDEPENDENT DATASET TO VALIDATE THE K VALUE. HISTORICALLY, THE OPTIMAL K FOR MOST DATASETS HAS BEEN BETWEEN 3-10. THAT PRODUCES MUCH BETTER RESULTS THAN 1NN.

Hamming Distance

$$D_H = \sum_{i=1}^{k} |x_i - y_i|$$

$$x = y \Rightarrow D = 0$$

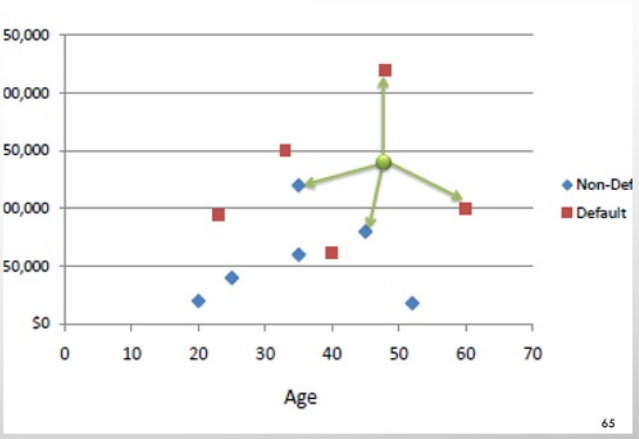
$$x \neq y \Rightarrow D = 1$$

X	Y	Distance
Male	Male	0
Male	Female	1

64

EXAMPLE

- CONSIDER THE FOLLOWING DATA CONCERNING CREDIT DEFAULT. AGE AND LOAN ARE TWO NUMERICAL VARIABLES (PREDICTORS) AND DEFAULT IS THE TARGET.



EXAMPLE

- WE CAN NOW USE THE TRAINING SET TO CLASSIFY AN UNKNOWN CASE (AGE=48 AND LOAN=$142,000) USING EUCLIDEAN DISTANCE. IF K=1 THEN THE NEAREST NEIGHBOR IS THE LAST CASE IN THE TRAINING SET WITH DEFAULT=Y.
- D = SQRT[(48-33)^2 + (142000-150000)^2] = 8000.01 >> DEFAULT=Y

Age	Loan	Default	Distance	
25	$40,000	N	102000	
35	$60,000	N	82000	
45	$80,000	N	62000	
20	$20,000	N	122000	
35	$120,000	N	22000	2
52	$18,000	N	124000	
23	$95,000	Y	47000	
40	$62,000	Y	80000	
60	$100,000	Y	42000	3
48	$220,000	Y	78000	
33	$150,000	Y	8000	1
48	**$142,000**	**?**		

Euclidean Distance

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

- With K=3, there are two Default=Y and one Default=N out of three closest neighbors. The prediction for the unknown case is again Default=Y.

66

EXAMPLE

STANDARDIZED DISTANCE

- ONE MAJOR DRAWBACK IN CALCULATING DISTANCE MEASURES DIRECTLY FROM THE TRAINING SET IS IN THE CASE WHERE VARIABLES HAVE DIFFERENT MEASUREMENT SCALES OR THERE IS A MIXTURE OF NUMERICAL AND CATEGORICAL VARIABLES.
- FOR EXAMPLE, IF ONE VARIABLE IS BASED ON ANNUAL INCOME IN DOLLARS, AND THE OTHER IS BASED ON AGE IN YEARS THEN INCOME WILL HAVE A MUCH HIGHER INFLUENCE ON THE DISTANCE CALCULATED. ONE SOLUTION IS TO STANDARDIZE THE TRAINING SET AS SHOWN ASIDE.
- USING THE STANDARDIZED DISTANCE ON THE SAME TRAINING SET, THE UNKNOWN CASE RETURNED A DIFFERENT NEIGHBOR WHICH IS NOT A GOOD SIGN OF ROBUSTNESS.

Age	Loan	Default	Distance
0.125	0.11	N	0.7652
0.375	0.21	N	0.5200
0.625	0.31	N	0.3160
0	0.01	N	0.9245
0.375	0.50	N	0.3428
0.8	0.00	N	0.6220
0.075	0.38	Y	0.6669
0.5	0.22	Y	0.4437
1	0.41	Y	0.3650
0.7	1.00	Y	0.3861
0.325	0.65	Y	0.3771
0.7	**0.61**	**?**	

Standardized Variable

$$X_s = \frac{X - Min}{Max - Min}$$

67

REFERENCES

T1: STEPHAN MARSLAND, **MACHINE LEARNING, AN ALGORITHMIC PERSPECTIVE**, CRC PRESS SECOND EDITION, 2015.

T2: ETHEM ALPAYDIN, **INTRODUCTION TO MACHINE LEARNING**, 2ND ED., PHI LEARNING PVT. LTD., 2013

R1: TOM M. MITCHELL, **MACHINE LEARNING**, MCGRAW-HILL EDUCATION (INDIAN EDITION), 2013