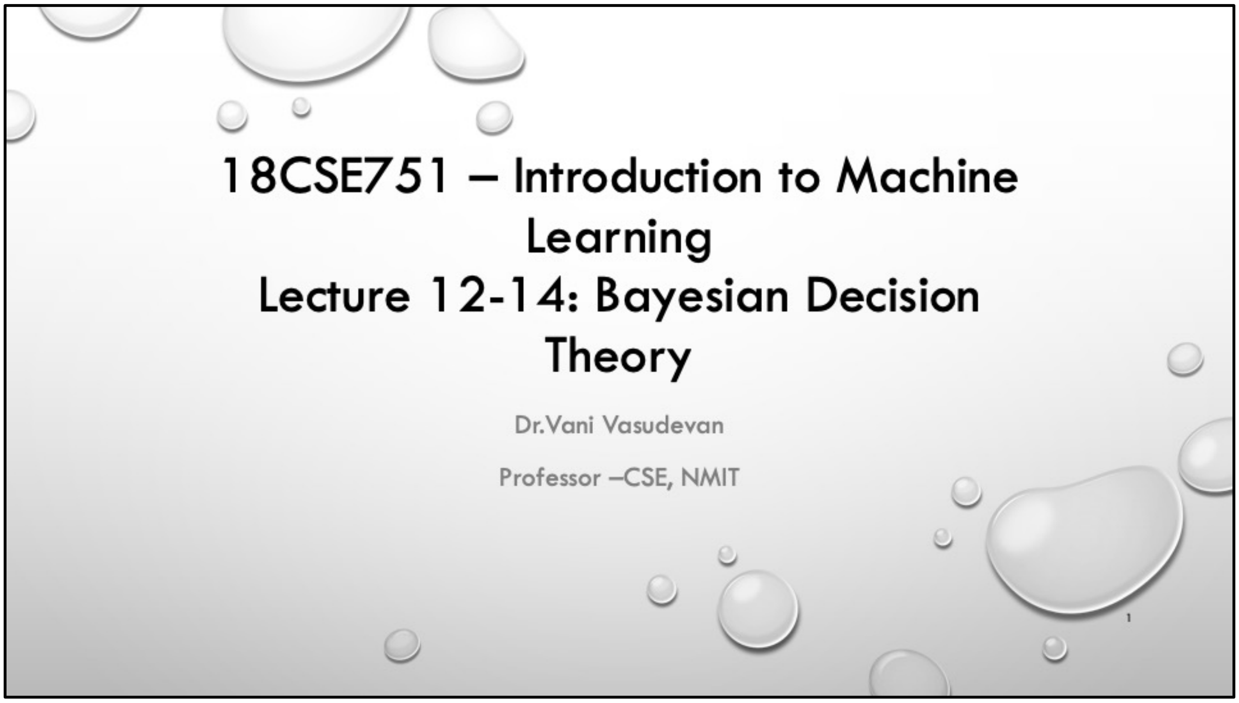
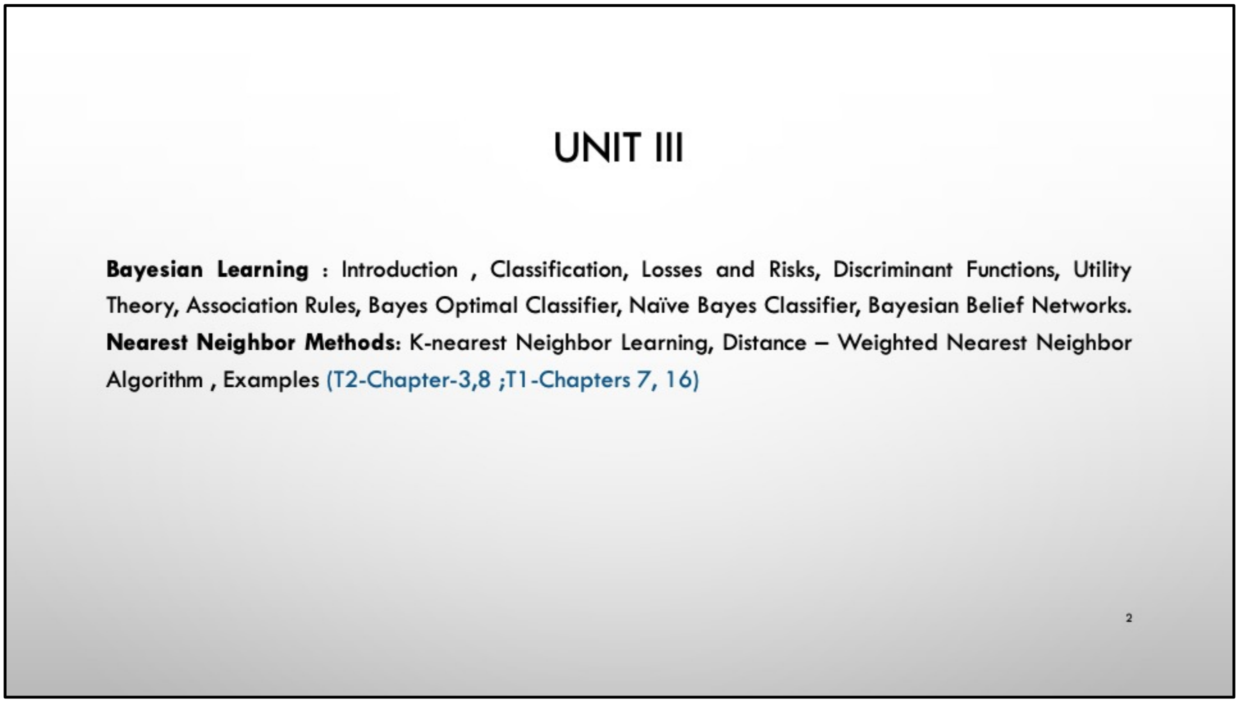


18CSE751 – Introduction to Machine Learning
Lecture 12-14: Bayesian Decision Theory

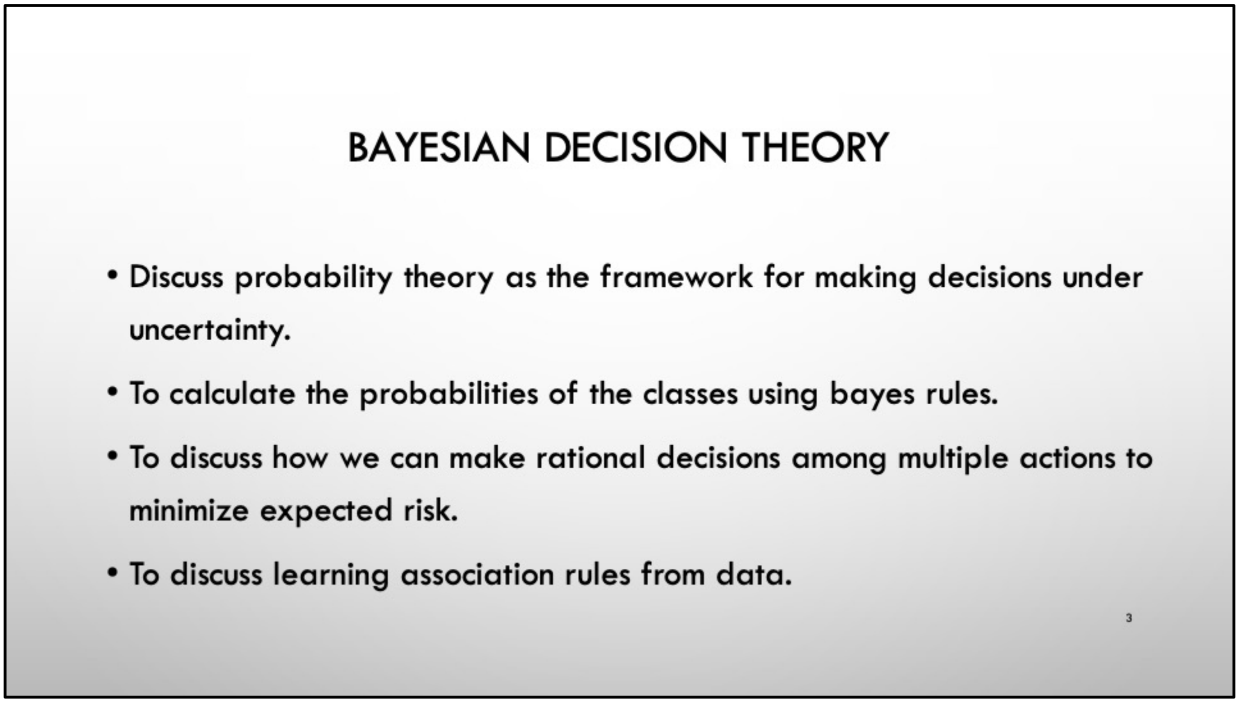
Dr.Vani Vasudevan

Professor –CSE, NMIT



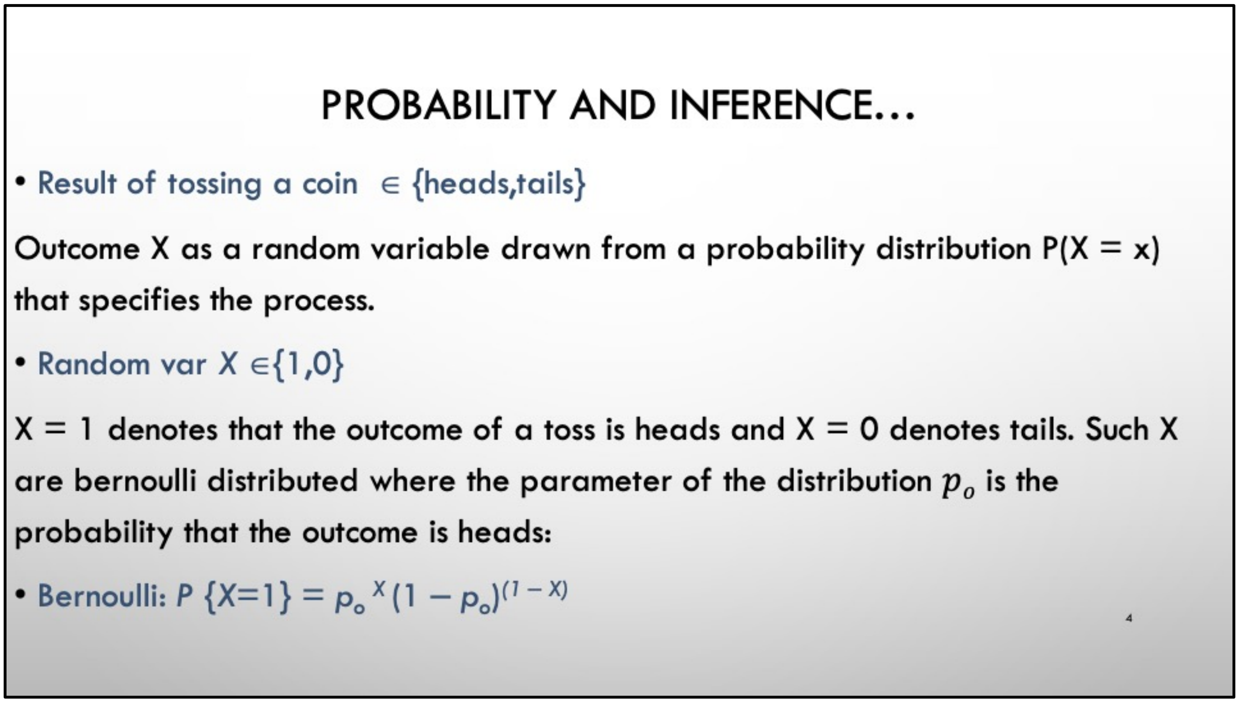
UNIT III

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)



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.



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.

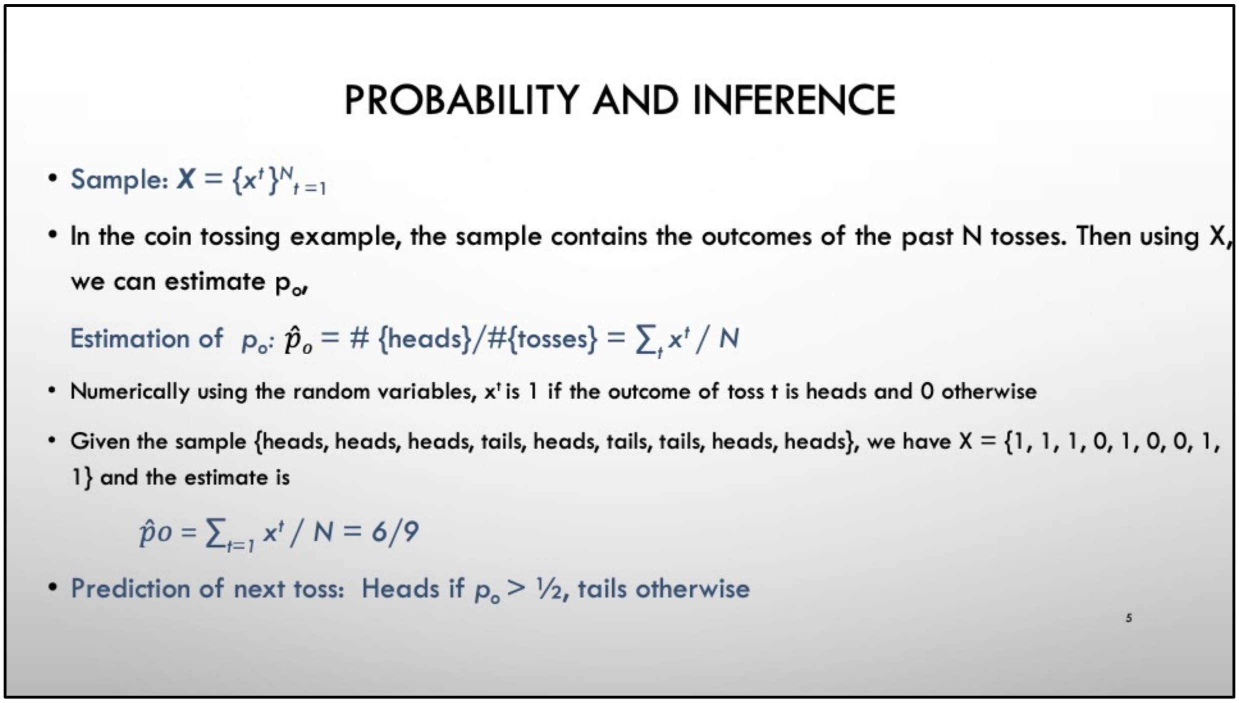
- Random var $X \in \{1,0\}$

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

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: $\boldsymbol{X} = \{x^t\}^N_{t=1}$
- 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: $\hat{p}_o = \#\{\text{heads}\}/\#\{\text{tosses}\} = \sum_t x^t / N$

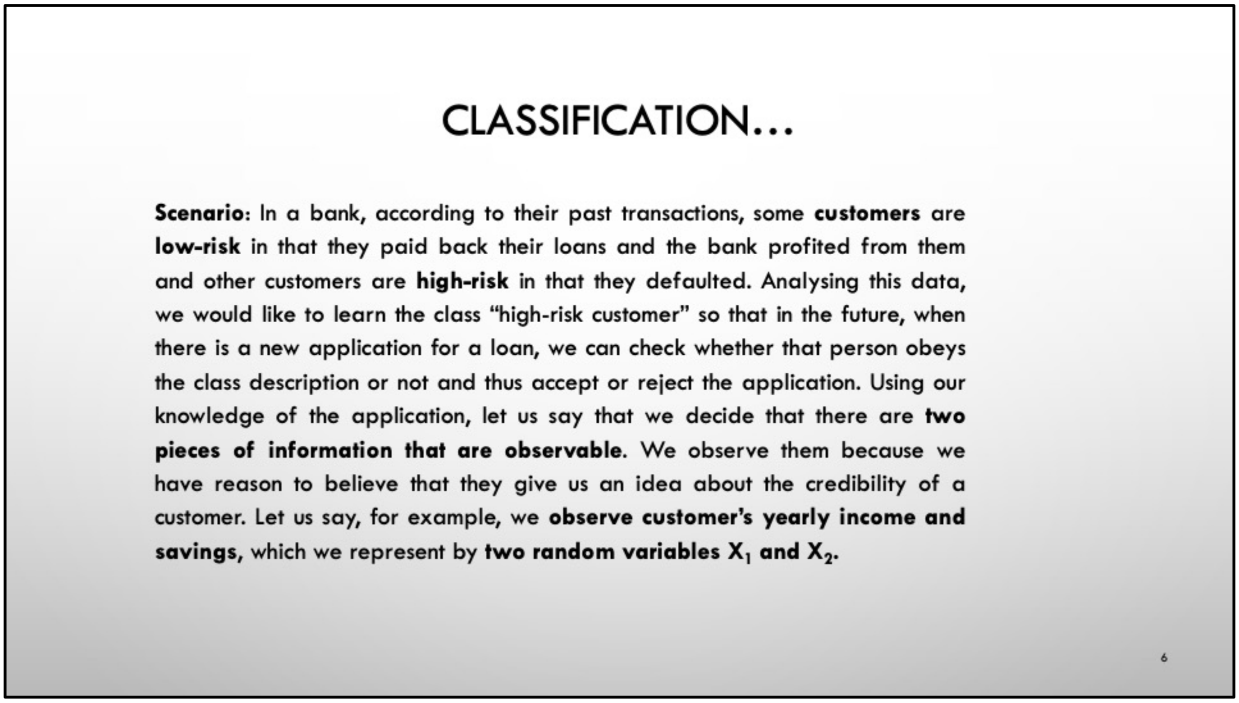
- Numerically using the random variables, x^t is 1 if the outcome of toss t is heads and 0 otherwise
- Given the sample {heads, heads, heads, tails, heads, tails, tails, heads, heads}, we have X = {1, 1, 1, 0, 1, 0, 0, 1, 1} and the estimate is

  $$\hat{p}o = \sum_{t=1} x^t / N = 6/9$$

- Prediction of next toss: Heads if $p_o > ½$, tails otherwise

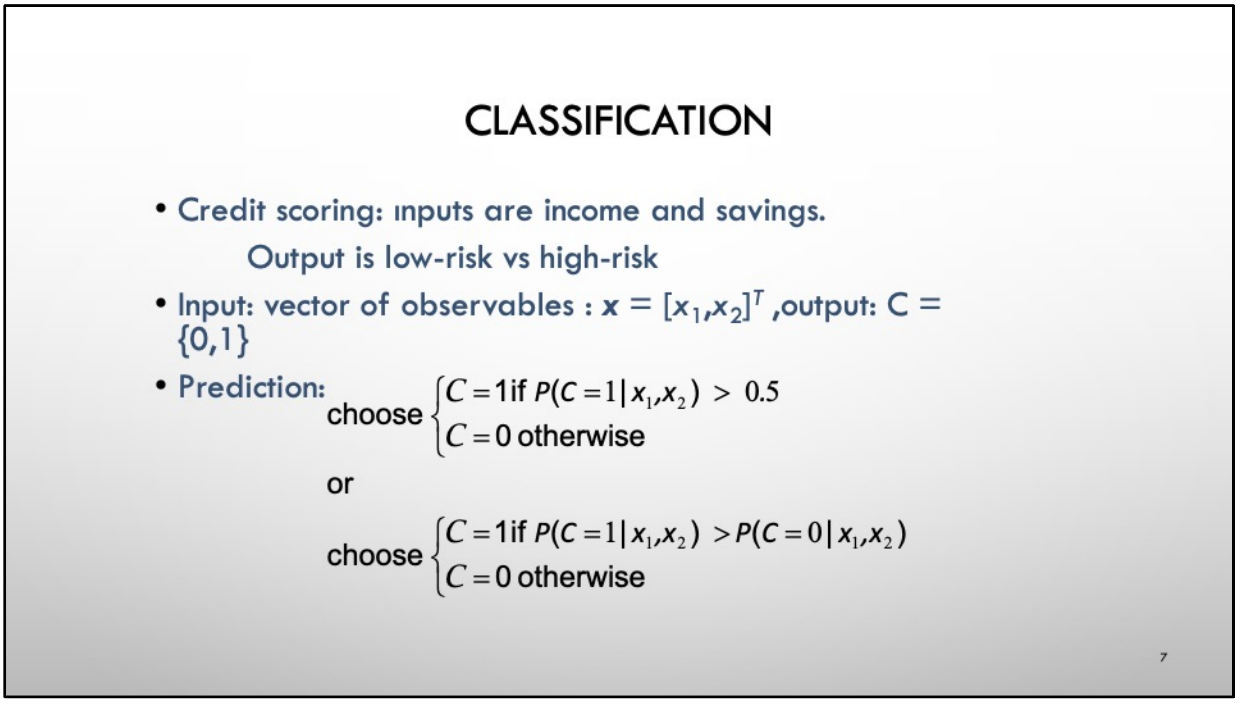
5

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: ınputs are income and savings.

  Output is low-risk vs high-risk
- Input: vector of observables : $\boldsymbol{x} = [x_1,x_2]^T$,output: C = {0,1}
- Prediction:

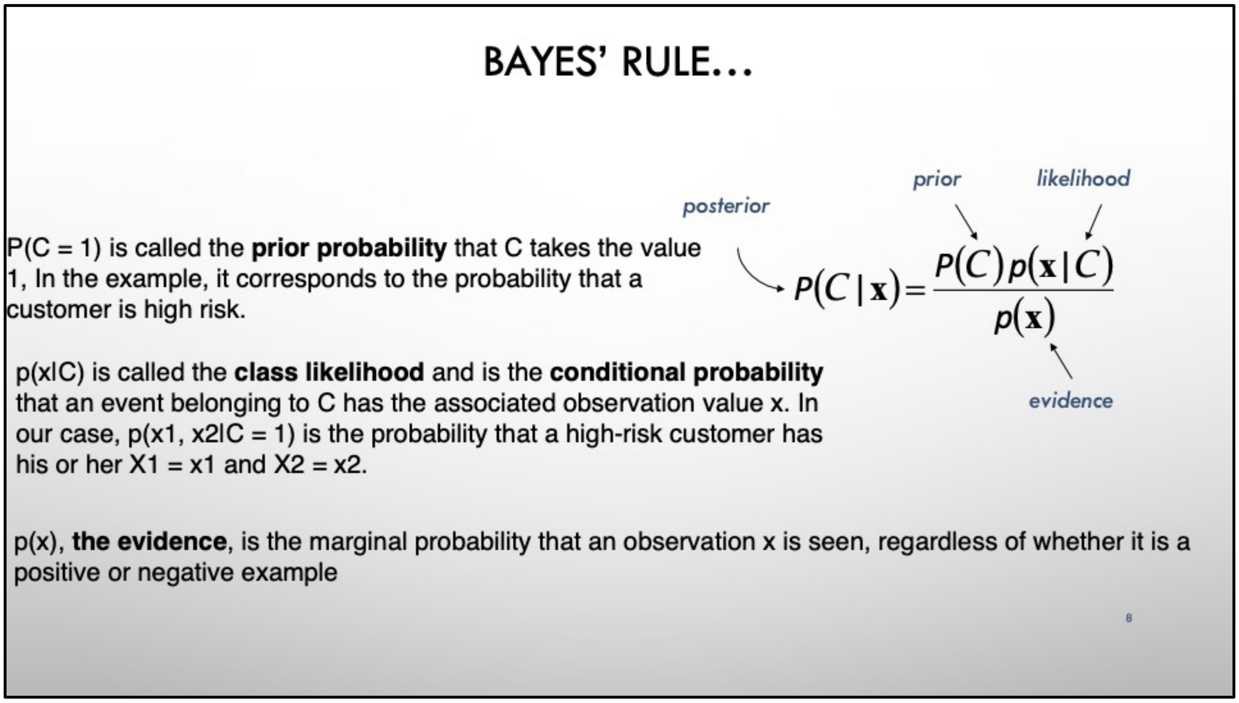
$$\text{choose} \begin{cases} C = 1 \text{ if } P(C = 1 \mid x_1,x_2) \; > \; 0.5 \\ C = 0 \text{ otherwise} \end{cases}$$

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))



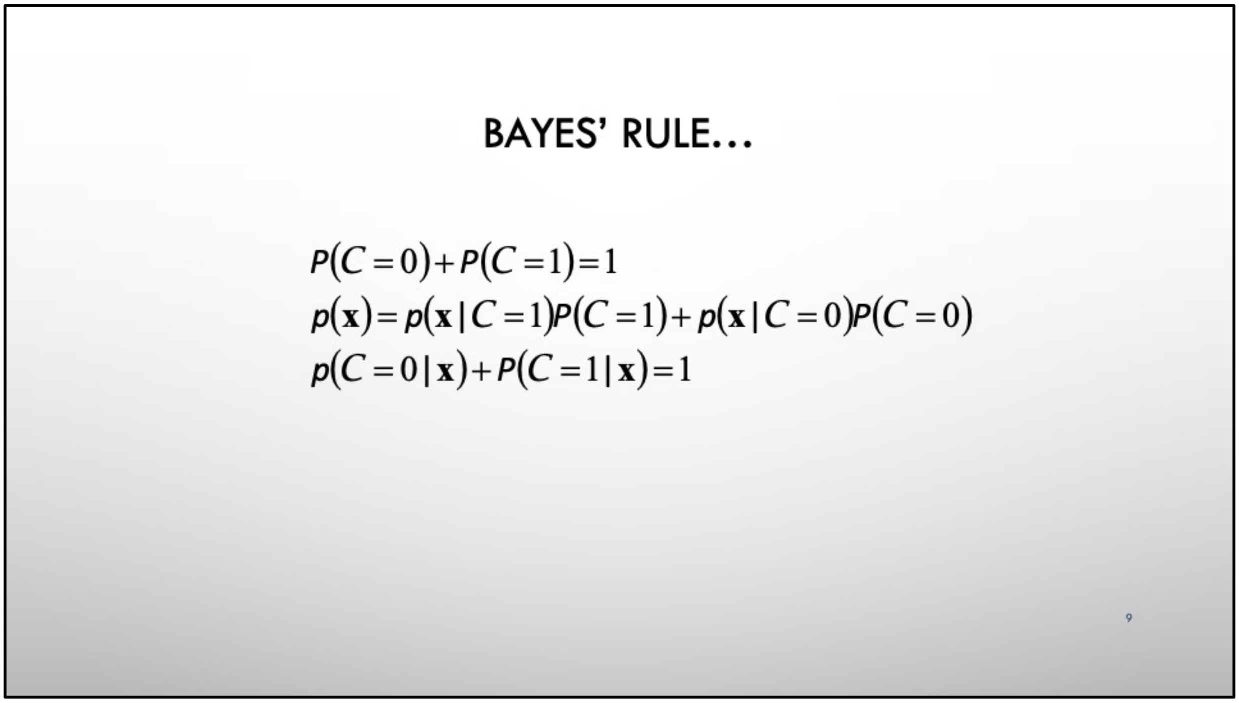
BAYES' RULE…

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.

$$\overset{\text{posterior}}{P(C \mid \mathbf{x})} = \frac{\overset{\text{prior}}{P(C)}\,\overset{\text{likelihood}}{p(\mathbf{x} \mid C)}}{\underset{\text{evidence}}{p(\mathbf{x})}}$$

p(x|C) 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, x2|C = 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



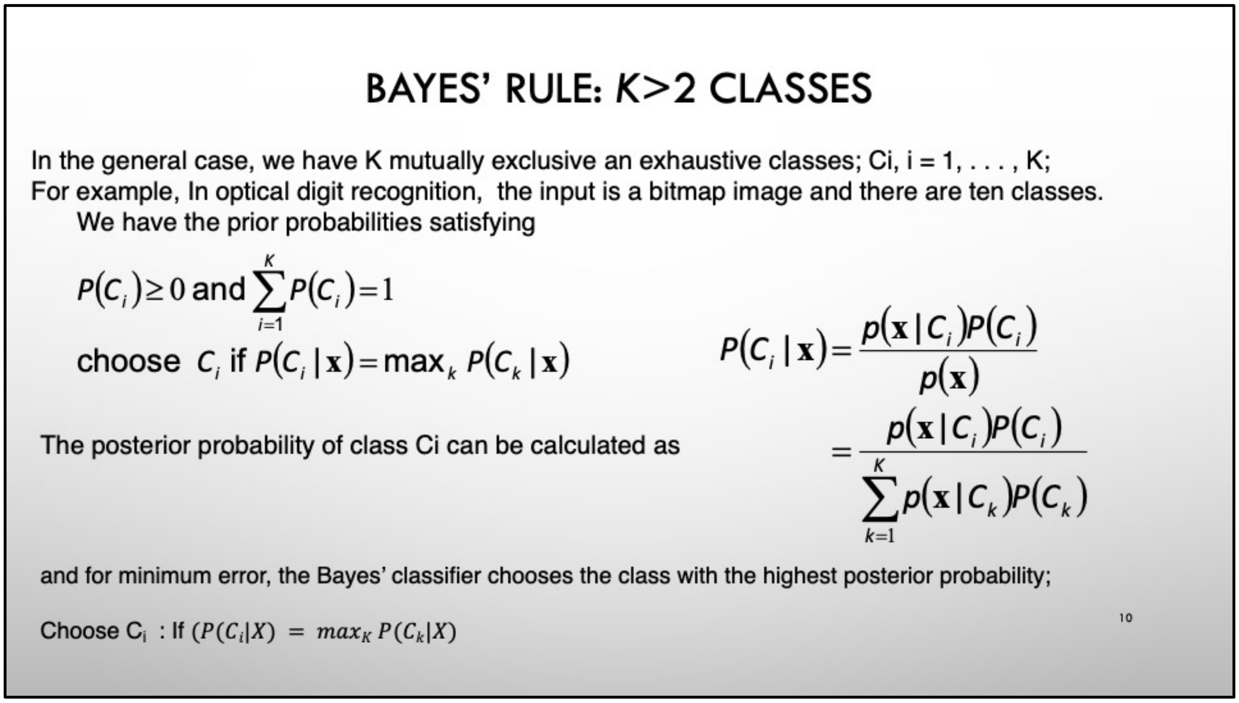
BAYES' RULE…

$$P(C=0)+P(C=1)=1$$

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

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

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

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

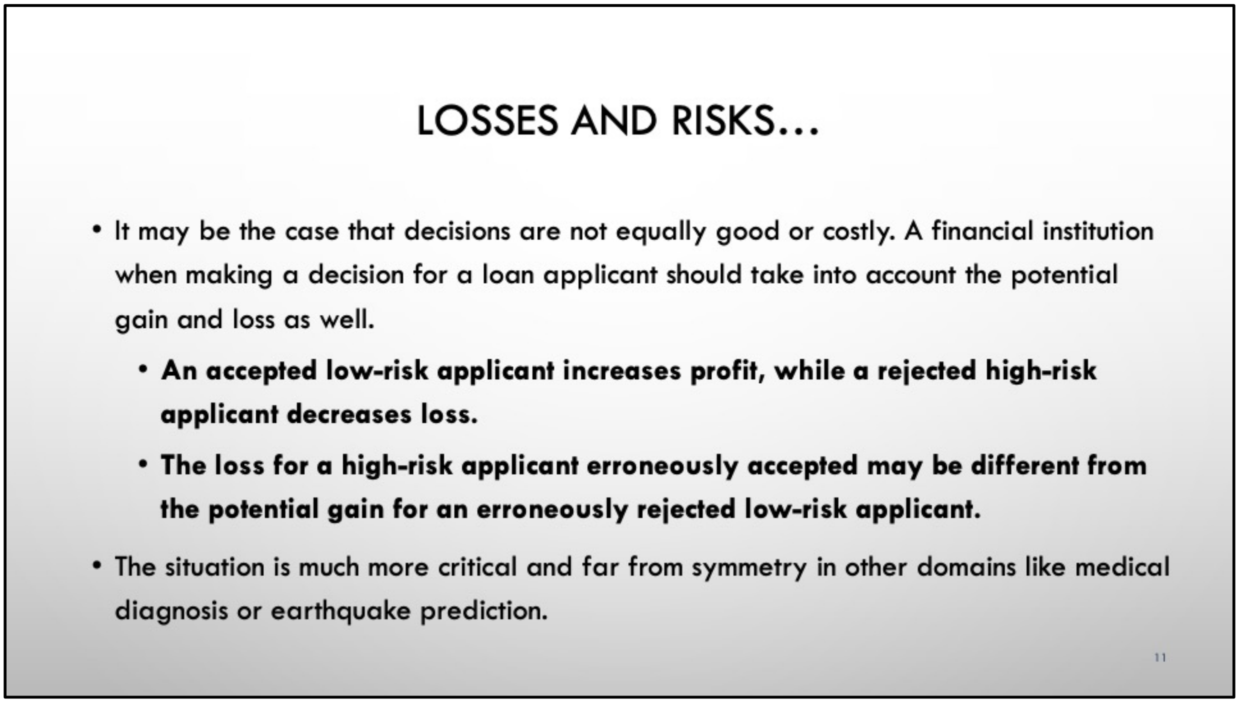
choose C_i if $P(C_i \mid \mathbf{x}) = \max_k P(C_k \mid \mathbf{x})$

The posterior probability of class Ci can be calculated as

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

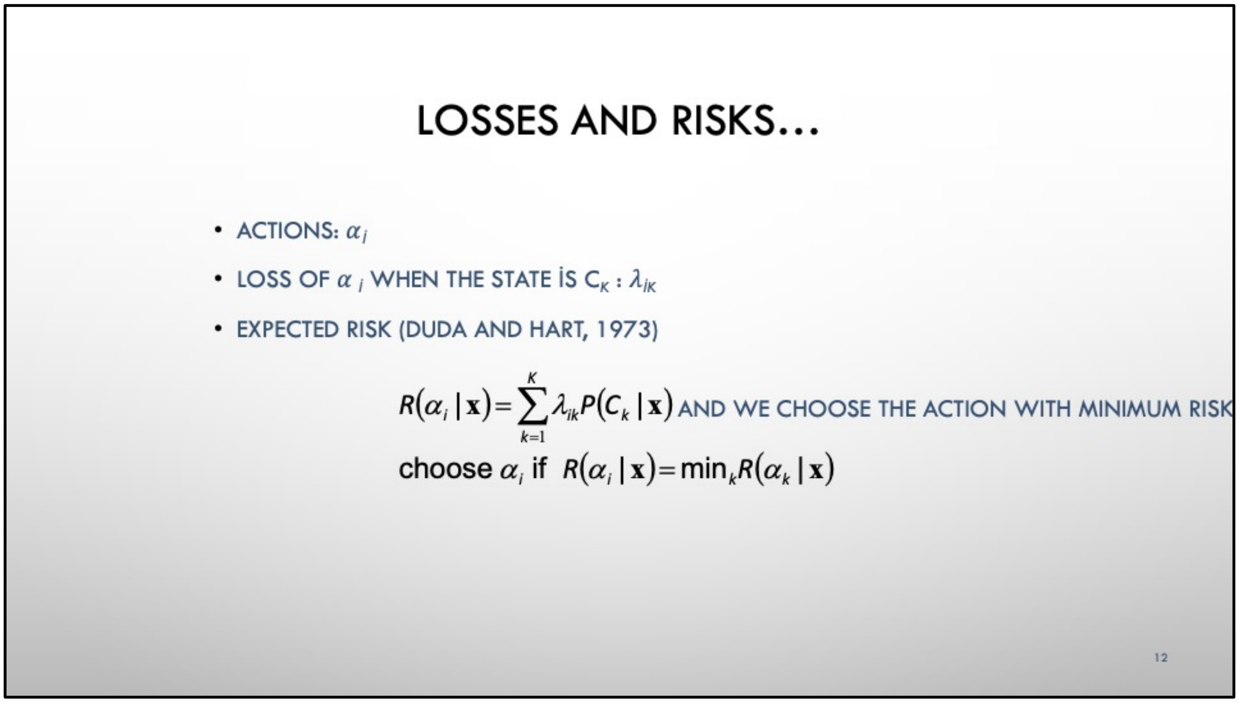
and for minimum error, the Bayes' classifier chooses the class with the highest posterior probability;

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



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.

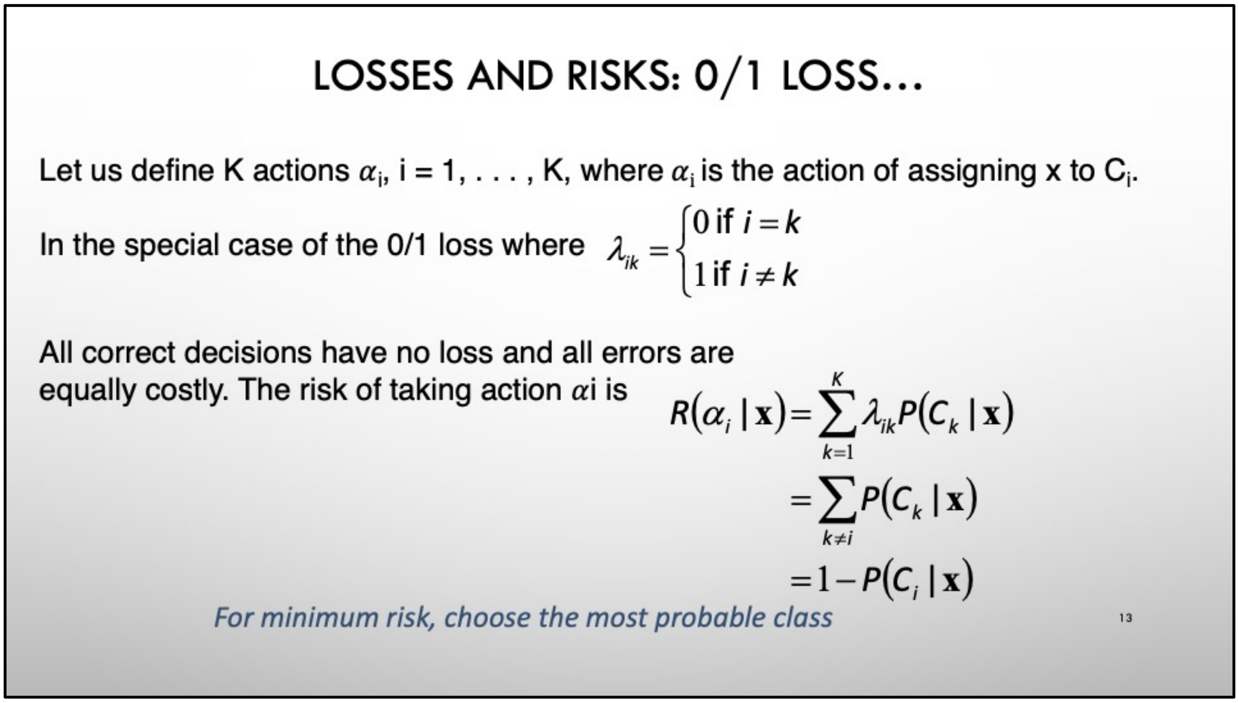


LOSSES AND RISKS…

- 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

choose α_i 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.

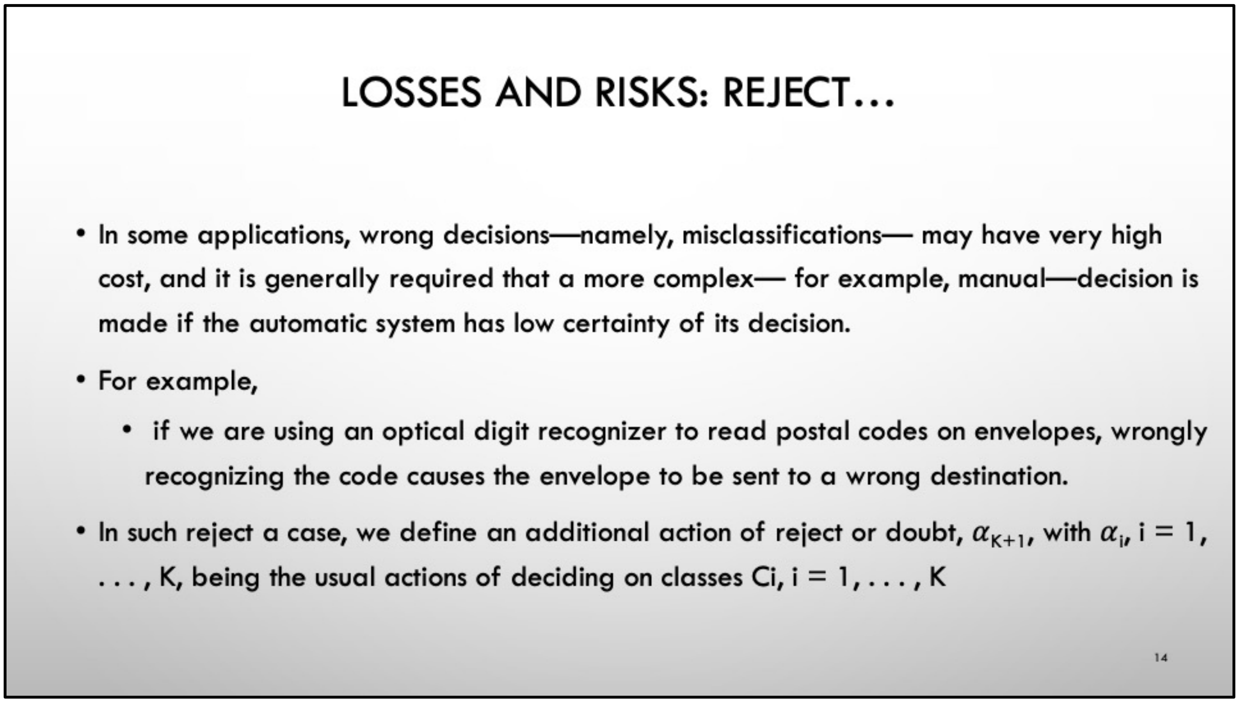
In the special case of the 0/1 loss where $\lambda_{ik} = \begin{cases} 0 \text{ if } i = k \\ 1 \text{ if } i \neq k \end{cases}$

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

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

For minimum risk, choose the most probable class

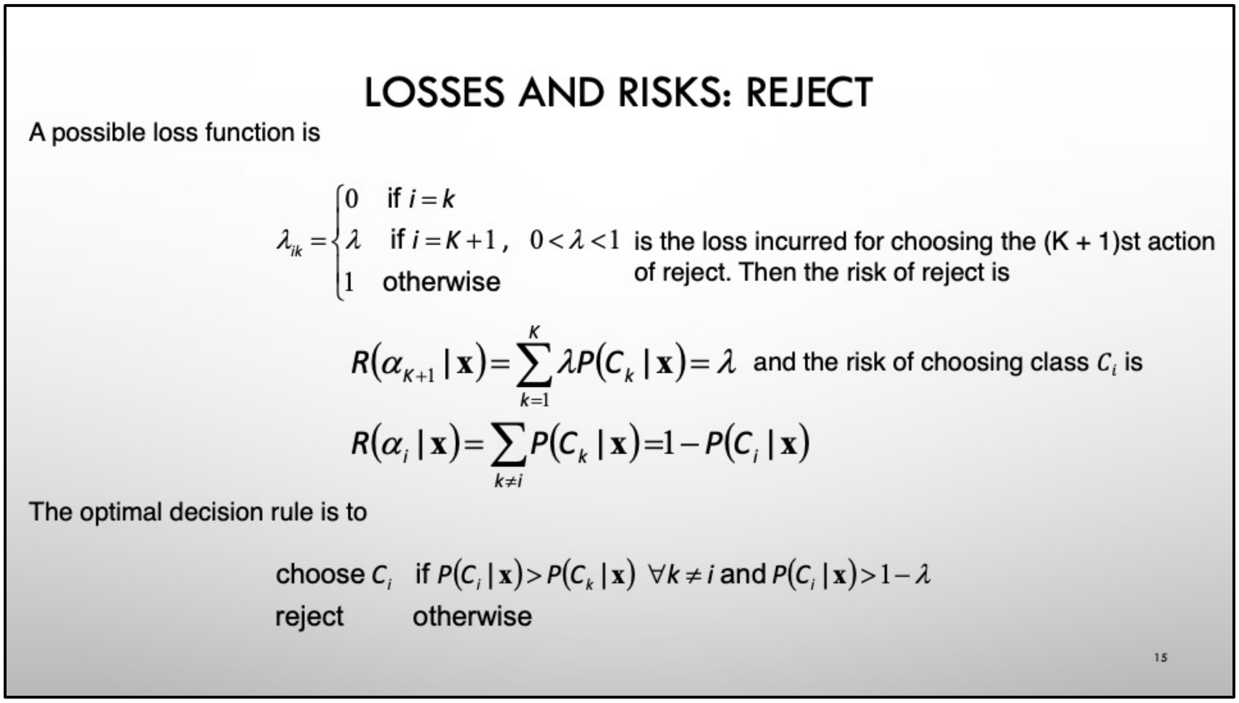
13



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

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

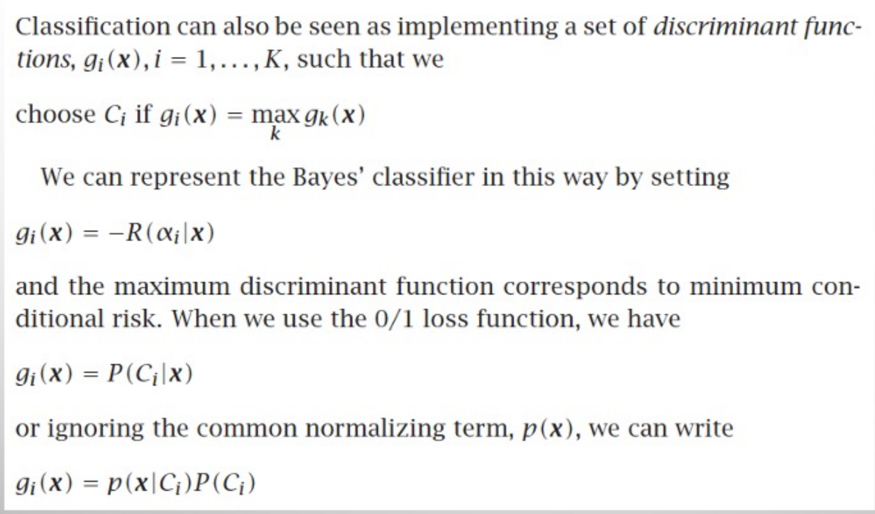
and the risk of choosing class C_i is

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

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

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

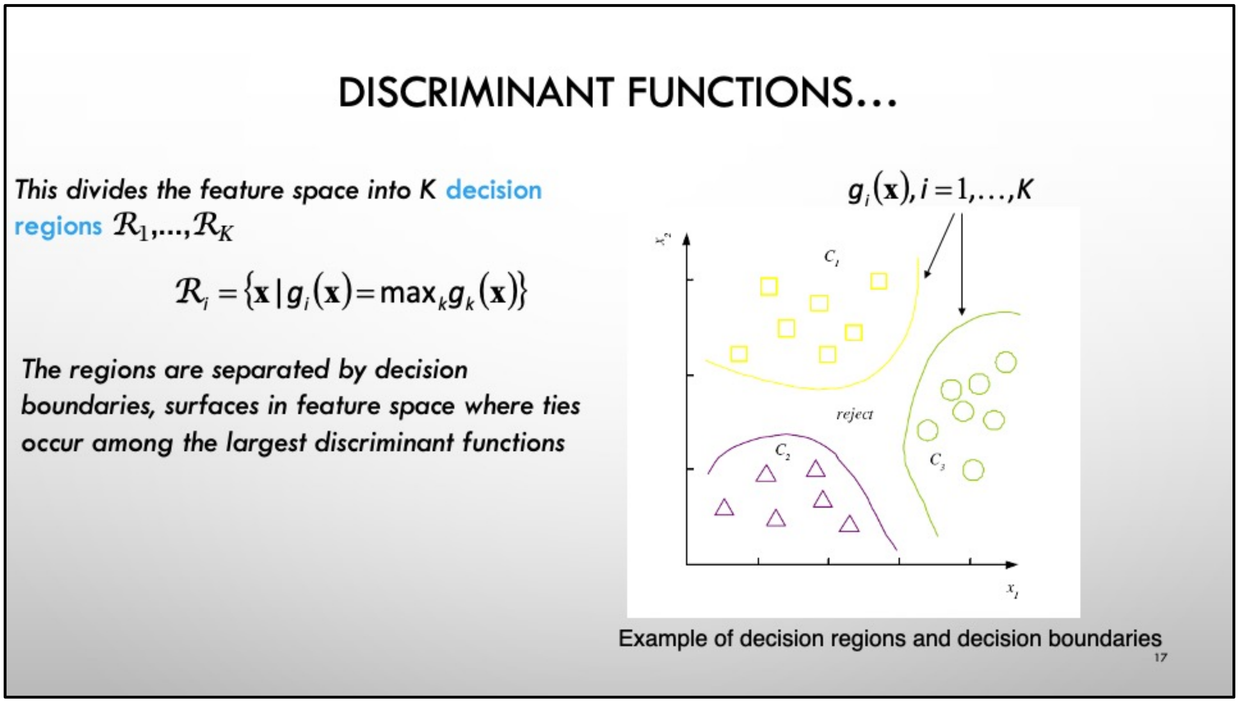
$g_i(\boldsymbol{x}) = -R(\alpha_i|\boldsymbol{x})$

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

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



DISCRIMINANT FUNCTIONS…

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

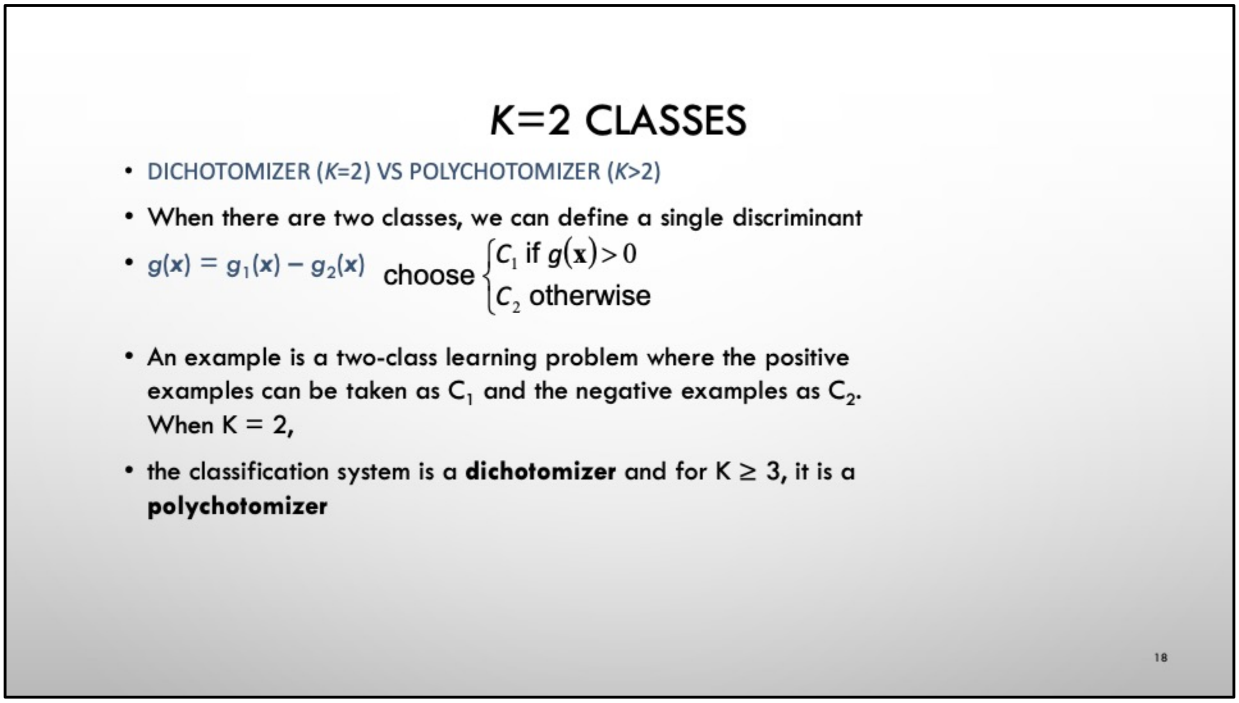
$$\mathcal{R}_i = \{\mathbf{x} \mid g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$

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

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

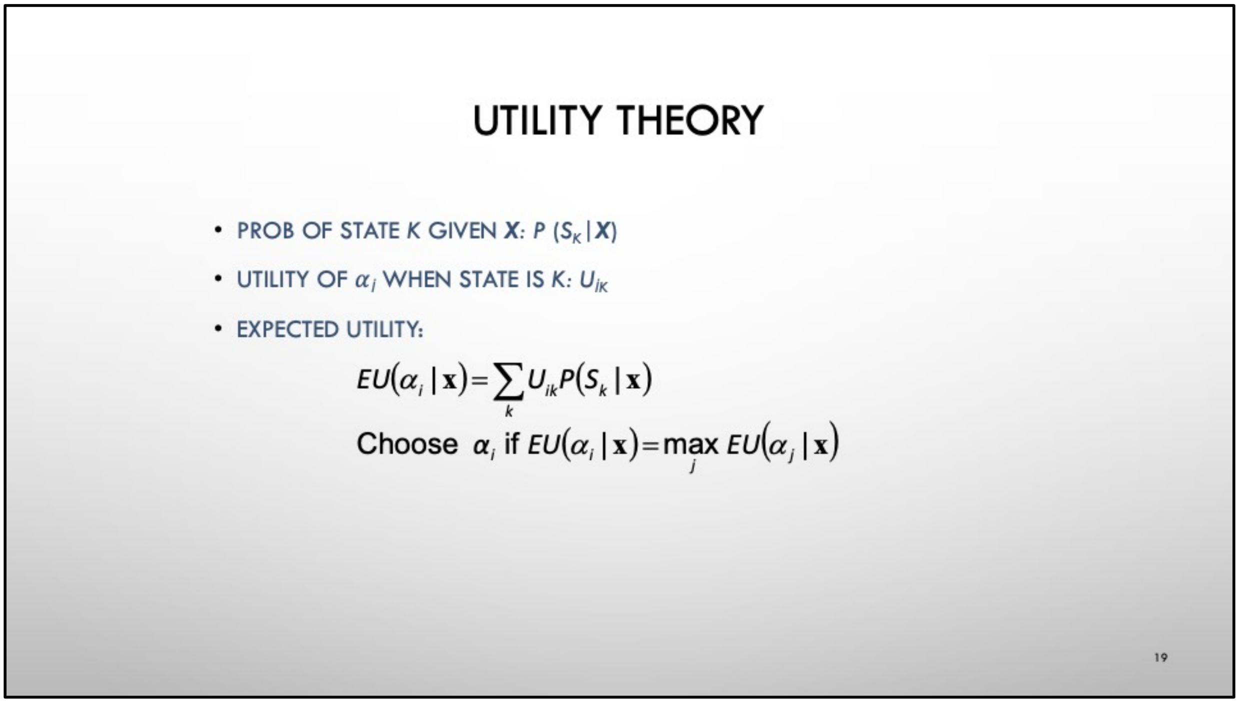
Example of decision regions and decision boundaries

17



K=2 CLASSES

- DICHOTOMIZER (K=2) VS POLYCHOTOMIZER (K>2)
- When there are two classes, we can define a single discriminant
- $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$ 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 ≥ 3, it is a **polychotomizer**

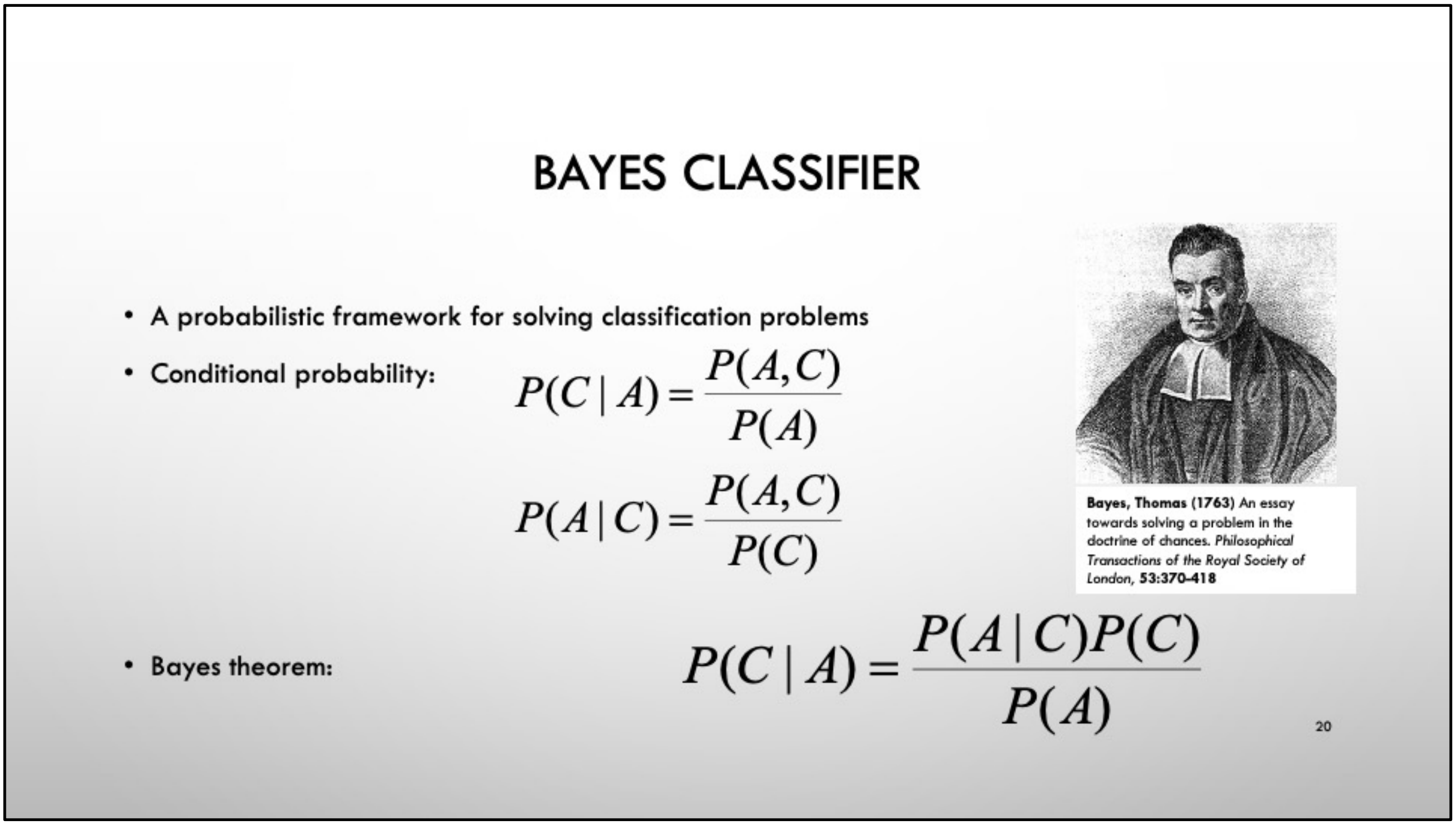


UTILITY THEORY

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

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

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



BAYES CLASSIFIER

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

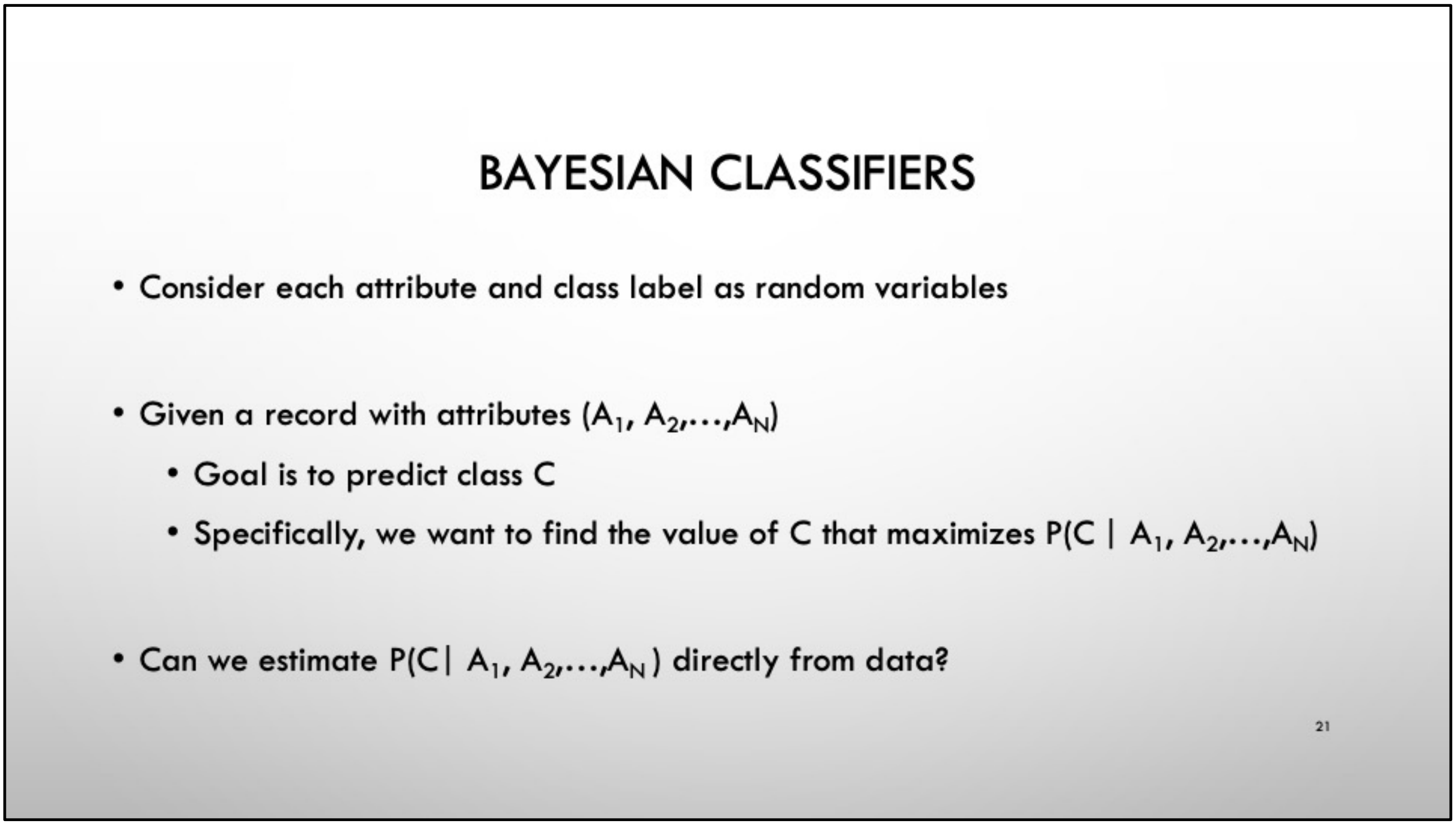
$$P(C \mid A) = \frac{P(A,C)}{P(A)}$$

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

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:

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

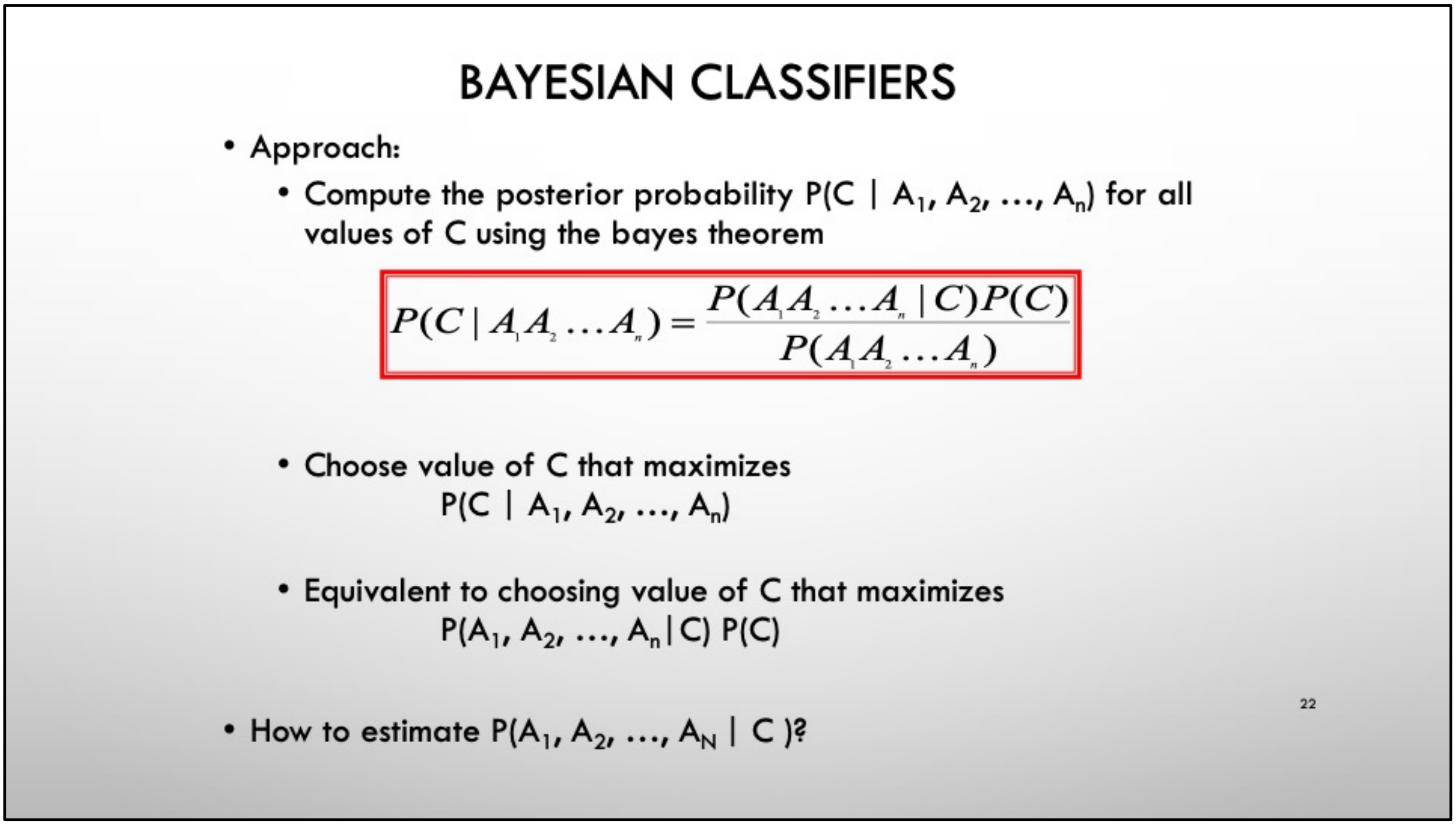


BAYESIAN CLASSIFIERS

- 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?



BAYESIAN CLASSIFIERS

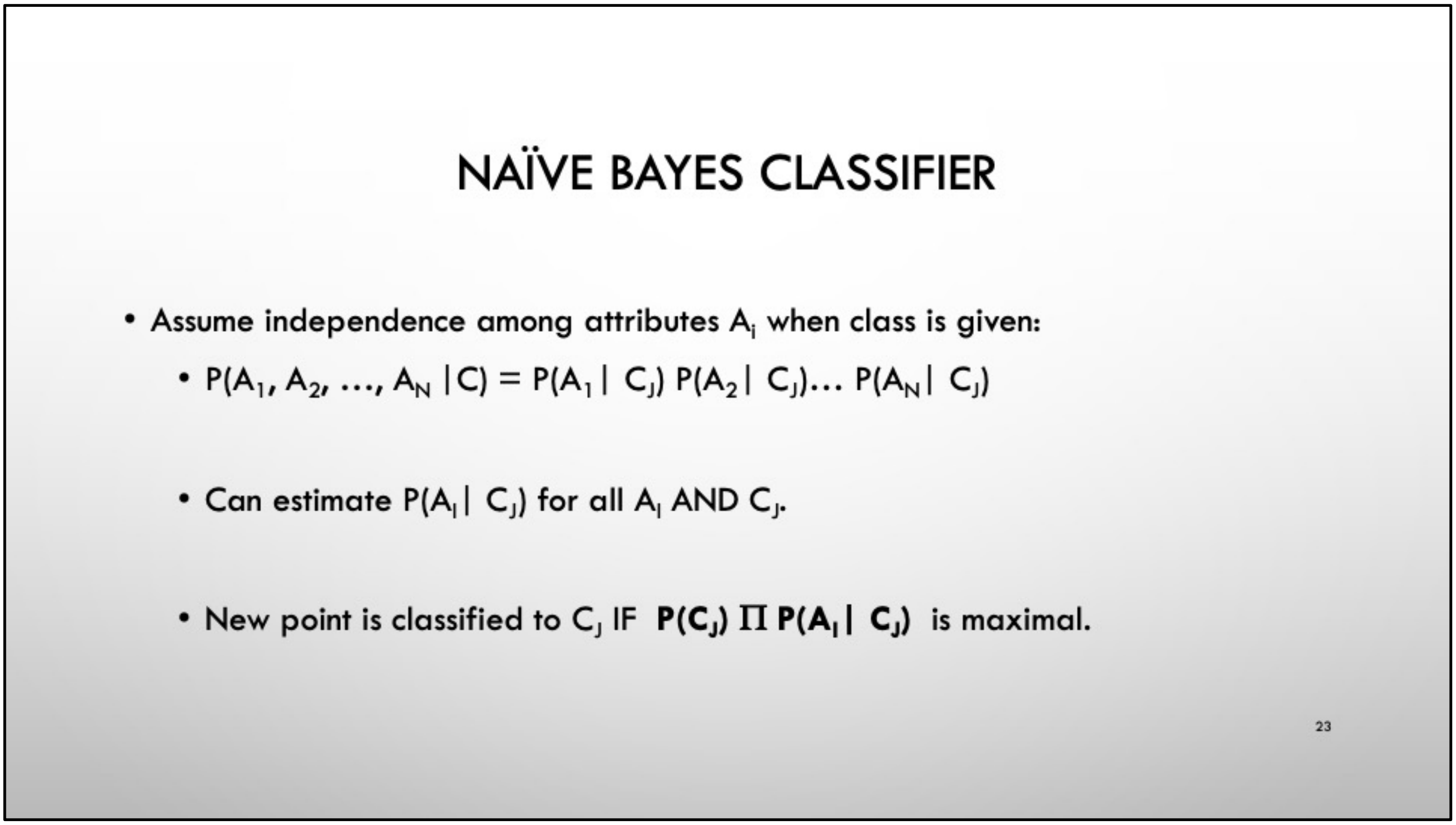
- Approach:
  - Compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the bayes theorem

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

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

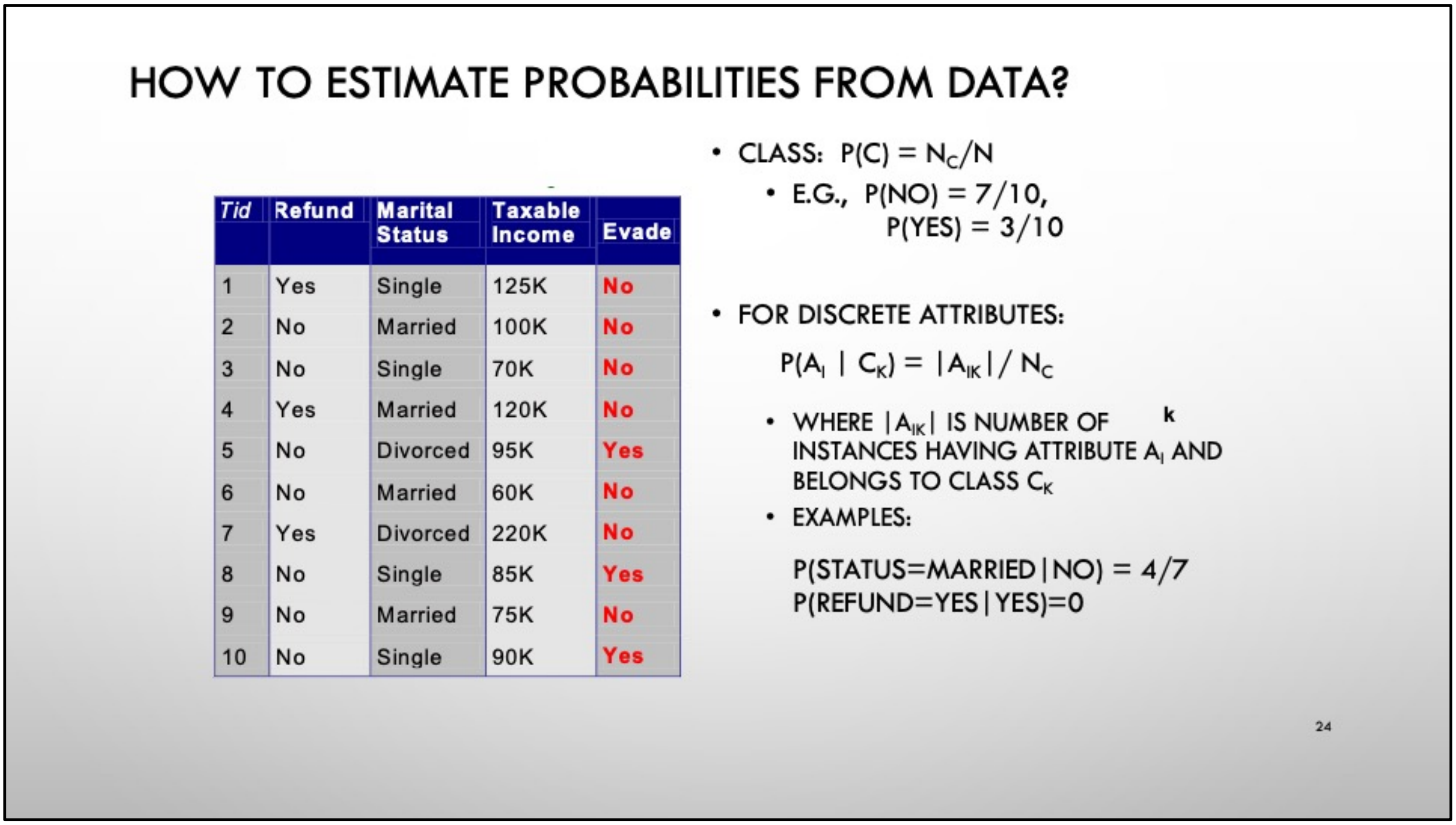
22

- 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 $\mathbf{P(C_J)\, \Pi\, P(A_I \mid C_J)}$ is maximal.



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

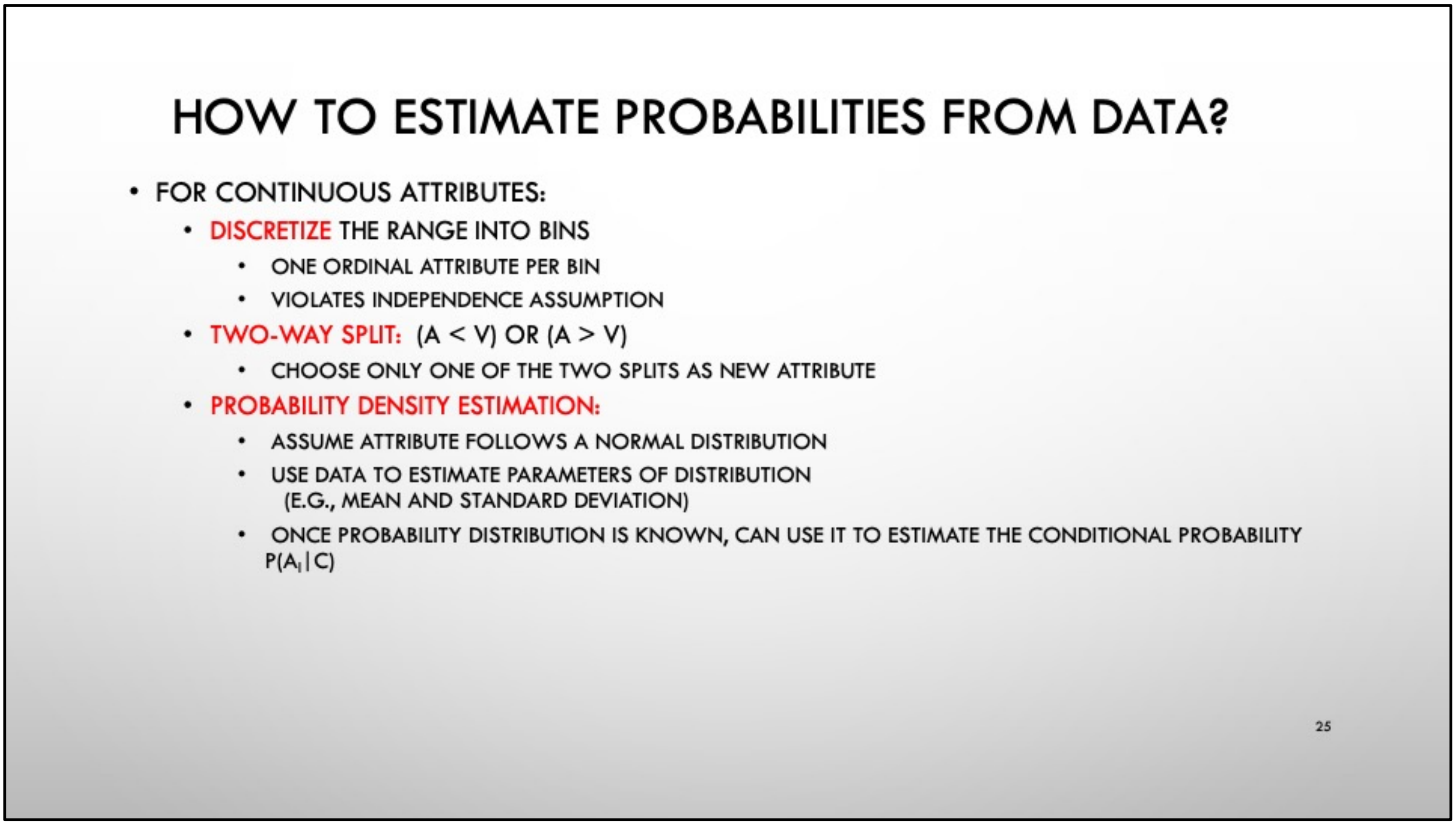
- CLASS: $P(C) = N_C/N$
  - E.G., $P(NO) = 7/10$,
    $P(YES) = 3/10$

- FOR DISCRETE ATTRIBUTES:

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

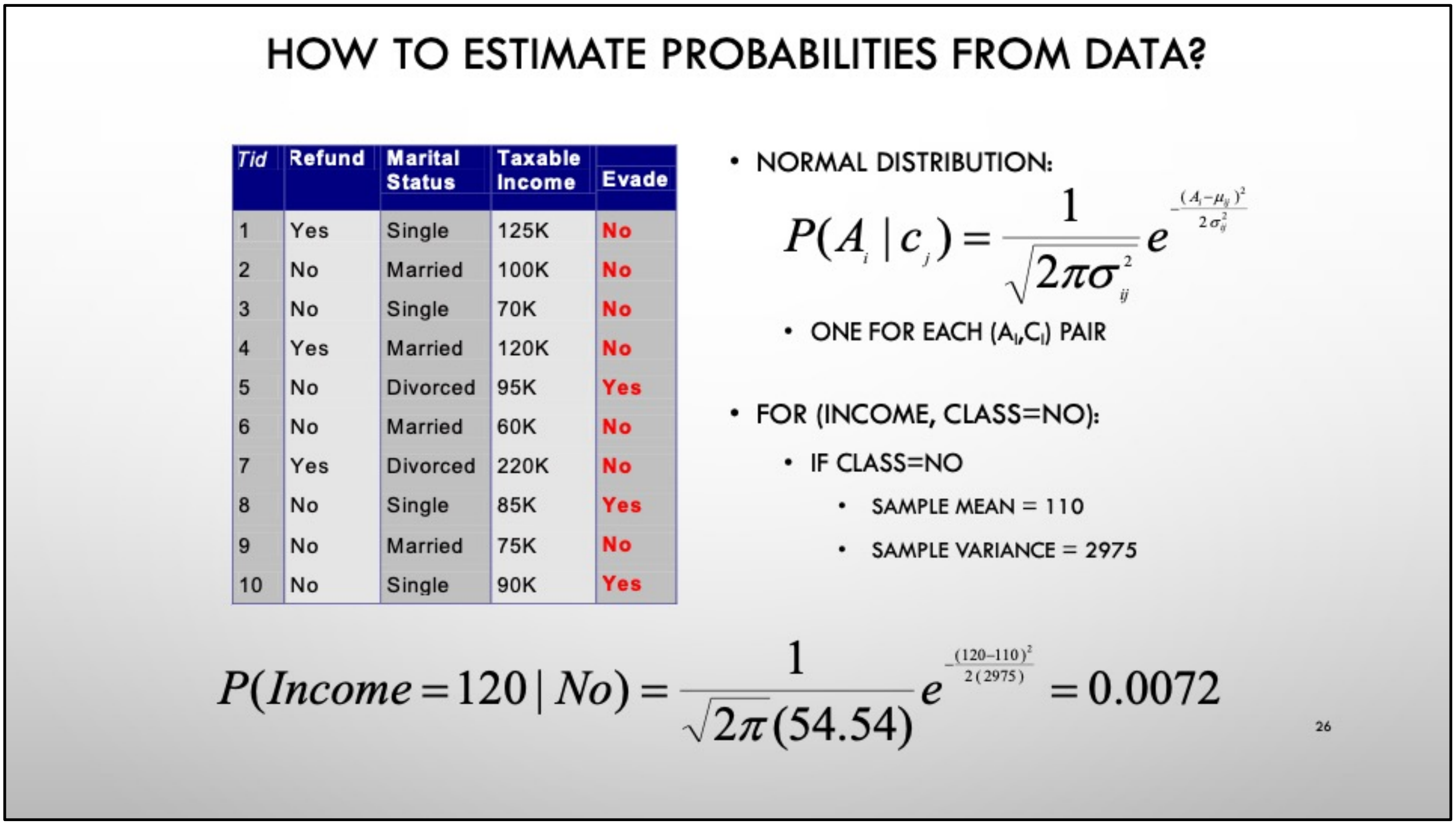
  - WHERE $|A_{IK}|$ IS NUMBER OF k
    INSTANCES HAVING ATTRIBUTE A_I AND BELONGS TO CLASS C_K
  - EXAMPLES:

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



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 $P(A_i|C)$



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

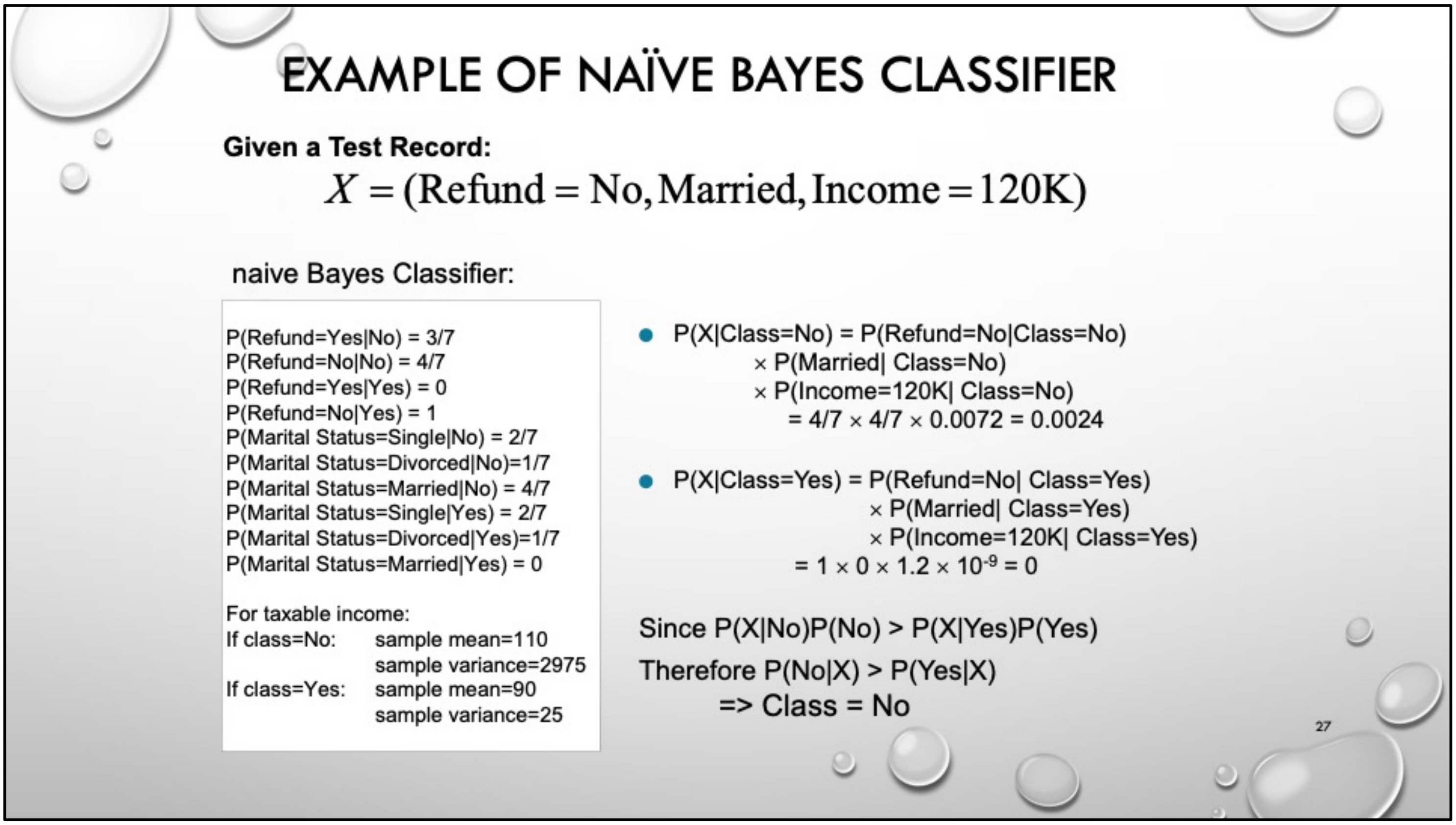
- NORMAL DISTRIBUTION:

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

  - ONE FOR EACH (A_i,C_i) PAIR

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

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



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:

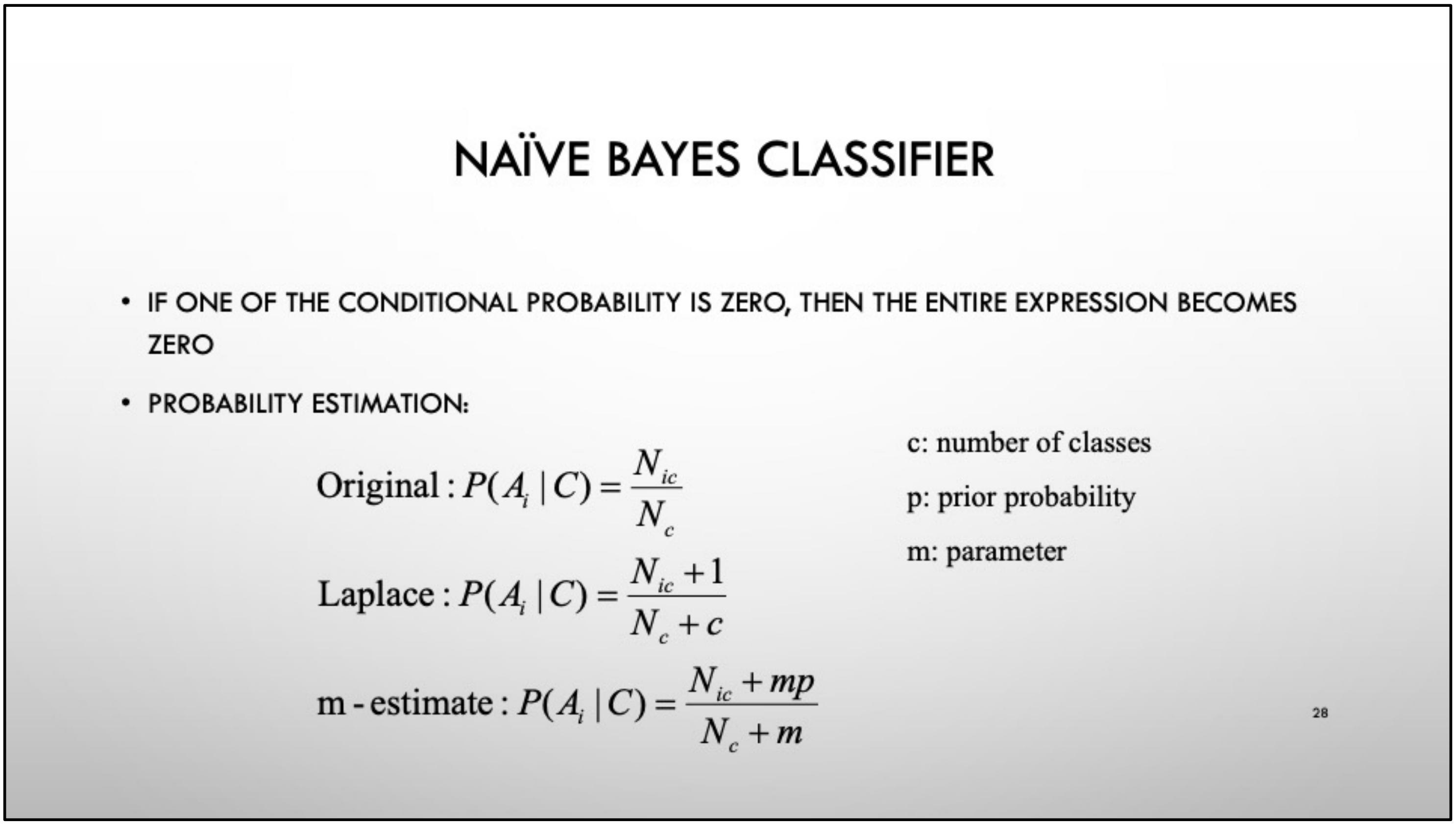
If class=No:	sample mean=110
	sample variance=2975
If class=Yes:	sample mean=90
	sample variance=25

- 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



NAÏVE BAYES CLASSIFIER

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

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

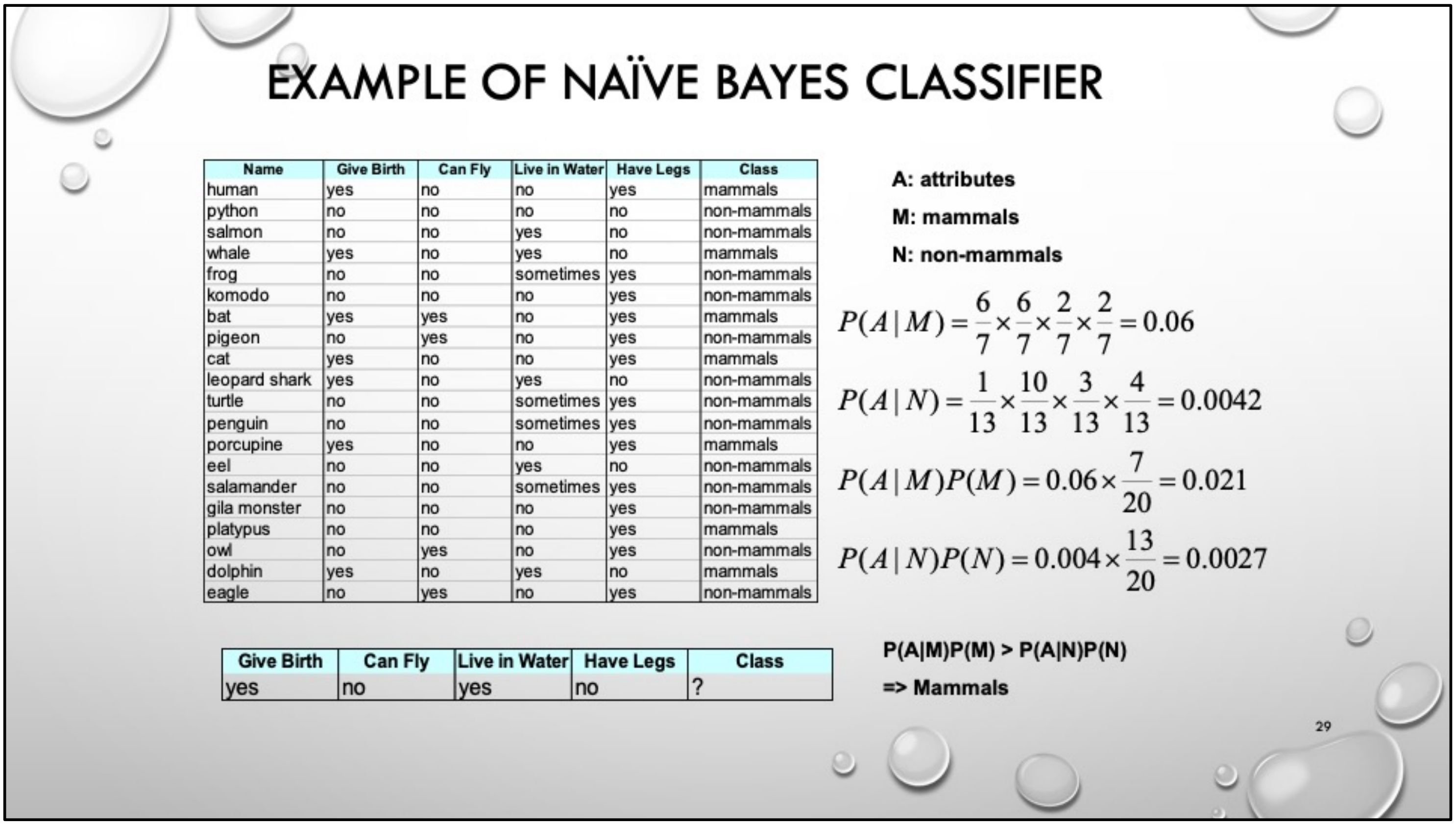
$$\text{Laplace} : P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

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

c: number of classes

p: prior probability

m: parameter



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

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

A: attributes

M: mammals

N: non-mammals

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

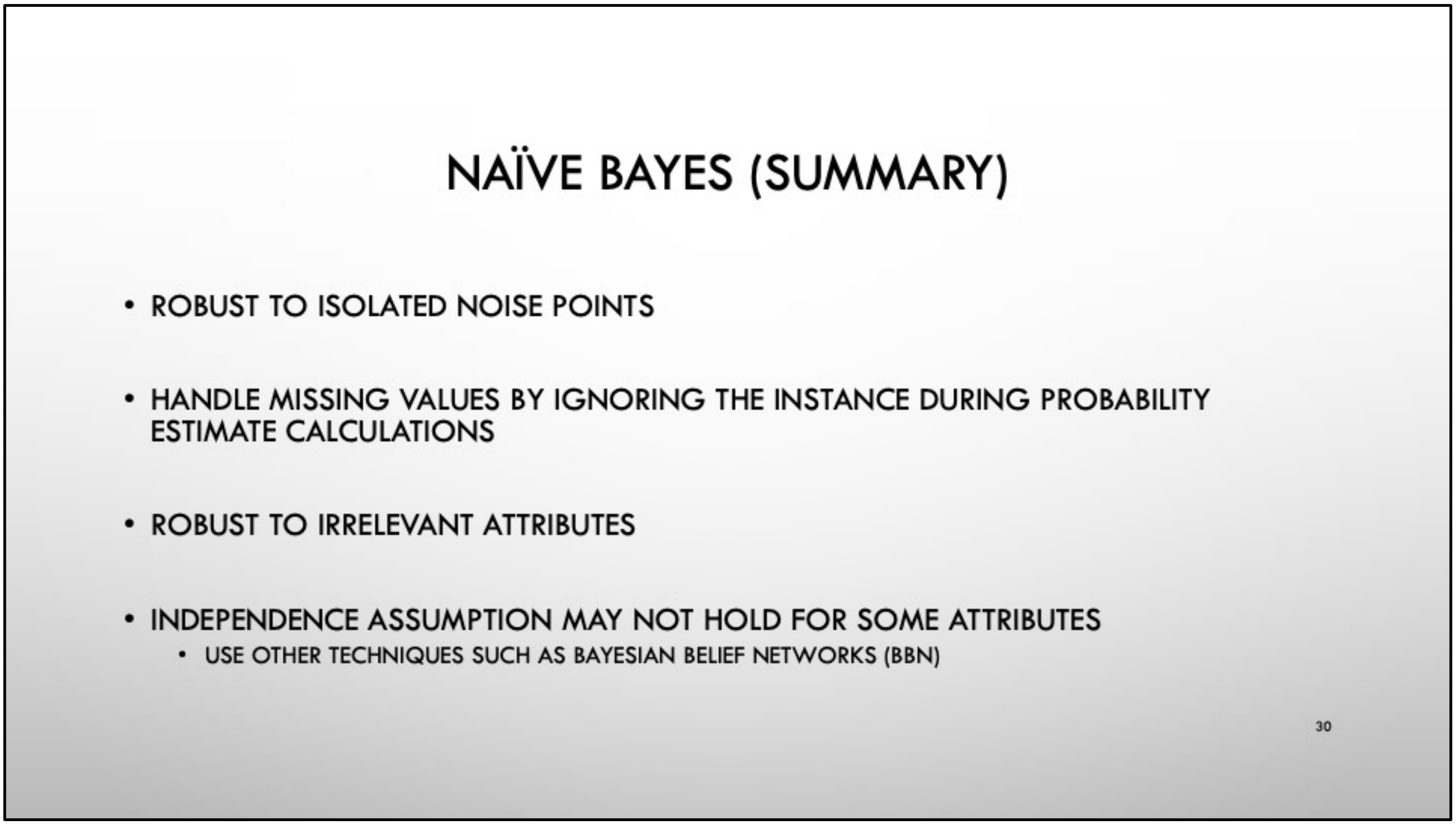
$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

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

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

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

=> Mammals



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)

30