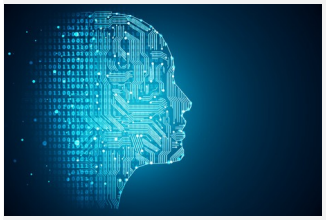


DS 310 Probabilistic Models

Vasant G. Honavar
Dorothy Foehr Huck and J. Lloyd Huck Chair in Biomedical Data Sciences and Artificial Intelligence
Professor of Data Sciences, Informatics, Computer Science and Engineering, Bioinformatics & Genomics, Public Health Sciences and Neuroscience
Director, Center for Artificial Intelligence Foundations and Scientific Applications
Associate Director, Institute for Computational and Data Sciences
Pennsylvania State University

vhonavar@psu.edu
http://faculty.ist.psu.edu/vhonavar
http://ailab.ist.psu.edu

Review: Probabilities and Probabilistic inference

- Given:
  - If Oksana studies, there is an 60% chance that she will pass the test; and a 40 percent chance that she will not.
  - If she does not study, there is 20% percent chance that she will pass the test and 80% chance that she will not.
- Observation: Oksana did not study.
- Example Inference task:
  - What is the chance that Oksana will pass the test?
  - What is the chance that she will fail?
- Probability theory generalizes propositional logic
  - Probability theory associates probabilities that lie in the interval [0,1] as opposed to 0 or 1 (exclusively)

Sources of uncertainty

Uncertainty modeled by Probabilistic assertions may

- In a deterministic world be due to
  - Laziness: failure to enumerate exceptions, qualifications, etc. that may be too numerous to state explicitly
  - Sensory limitations
  - Ignorance: lack of relevant facts etc.
- In a stochastic world be due to
  - Inherent uncertainty (as in quantum physics)

The framework is agnostic about the source of uncertainty

bility Theory

eriment has a set of **potential outcomes**, e.g.,

e

another data sample

pace of a **random experiment** is the set of all
omes, e.g., {1, 2, 3, 4, 5, 6}

earning the sample spaces can be ***very*** large

N binary features, the sample space is 2^N

ility theory

et of the sample space

4, 6}

5}

3

eature has particular values

ple is described by the values of features

Events

We're interested in probabilities of events

- p({2})
- p(label=cancer)
- p(tumor_present = 1)
- p(smoker = 1)

riables

HHT	HTH	HTT	THH	THT	TTH	TTT
2	2	1	2	1	1	0

ble is a mapping from the sample space to a
events)

vhose values we want to measure in an experiment

ndom variable, *X*, could be the number of heads in
riment

Random variables

- We're interested in the probability of the different values of a random variable e.g., number of heads in 3 coin tosses
- The definition of probabilities over *all* of the possible values of a random variable defines a **probability distribution**

	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
X	3	2	2	1	2	1	1	0

$P(X = 3) = P(HHH) = \frac{1}{8}$

$P(X = 0) = P(TTT) = \frac{1}{8}$

$P(X = 2) = 1/8 + 1/8 + 1/8 = 3/8$

$P(X = 1) = 1/8 + 1/8 + 1/8 = 3/8$

Probability distribution

To be explicit

- A probability distribution assigns probability values to *all possible values* of a random variable
- These values must be ≥ 0 and ≤ 1
- These values must sum to 1 for all possible values of the random variable

Unconditional probability

Simplest form of probability is

- P(X)
  - What is the probability of heads?
  - What is the probability of surviving cancer?
  - What is the probability of a car review containing the word “automatic”?
  - What is the probability of a passenger on the Titanic being under 21 years old?
  - …

Joint distribution

We can also talk about probability distributions over multiple variables

P(X,Y)

- probability of X *and* Y
- a distribution over the cross product of possible values

MLPass AND EngPass	P(MLPass, EngPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

Joint distribution

Still a probability distribution

- all values between 0 and 1, inclusive
- all values sum to 1

All questions/probabilities of the two variables can be calculate from the joint distribution

MLPass AND EngPass	P(MLPass, EngPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

What is P(ENGPass)?

Joint distribution

Still a probability distribution

- all values between 0 and 1, inclusive
- all values sum to 1

All questions/probabilities of the two variables can be calculate from the joint distribution

MLPass AND EngPass	P(MLPass,EngPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

P(EngPass)=0.84

How did we figure that out?

Joint distribution

$$P(x) = \sum_{y \in Y} p(x, y)$$

MLPass AND EngPass	P(MLPass, EngPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

EngPass	P(EngPass)
true	0.84
false	0.16

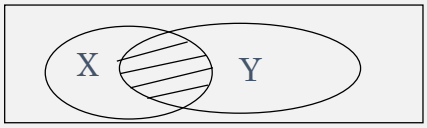
MLPass	P(MLPass)
true	0.81
false	0.19

Conditional probability

- As we learn more information, we can update our probability distribution
- $P(X|Y)$ (read "probability of X *given* Y") models this
  - What is the probability of a heads *given* that both sides of the coin are heads?
  - What is the probability the document is about cars, given that it contains the word "fuel"?
  - What is the probability of the word "golf" given that the sentence also contains the word "course"?
- Notice that $P(X|Y)$ is still a distribution over the values of X

Conditional probability

- $P(X|Y) = ?$



In terms of pior and joint distributions, what is the conditional probability distribution?

Conditional probability

$$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

Given that Y has occurred, in what proportion of those events does X also occur?

Conditional probability

$$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

MLPass AND EngPass	P(MLPass, EngPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

What is:
P(MLPass=true | EngPass=false)?

Conditional probability

MLPass AND EngPass	P(MLPass, EngPass)
true, true	.80
true, false	.01
false, true	.04
false, false	.15

$$p(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

What is:
p(MLPass=true | EngPass=false)?

$$\frac{P(\ MLPass = True, EngPass = False)}{P(EngPass = False)} = \frac{0.01}{0.15 + 0.01} = \frac{1}{16}$$

Notice this is very different than $P(MLPass = true) = 0.81$

Both are distributions over X

Unconditional/prior probability

Conditional probability

$p(X)$

MLPass	P(MLPass)
true	0.81
false	0.19

$p(X \mid Y)$

MLPass	P(MLPass | EngPass=false)
true	0.0625
false	0.9375

Probability as a measure over possible worlds

- Suppose I have two coins – one a normal fair coin, and the other with 2 heads.
- I pick a coin at *random* and toss it.
- What is the probability that the outcome is a head?

$$\Omega = \{(Fair, H), (Fair, T), (Rigged, H), (Rigged, T)\}$$

$$\mu = \left\{\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0\right\}$$

$$\Pr(H) = \sum_{\omega \models H} \mu(\omega) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$

Conditional probability as a Measure over Possible worlds not ruled out by evidence

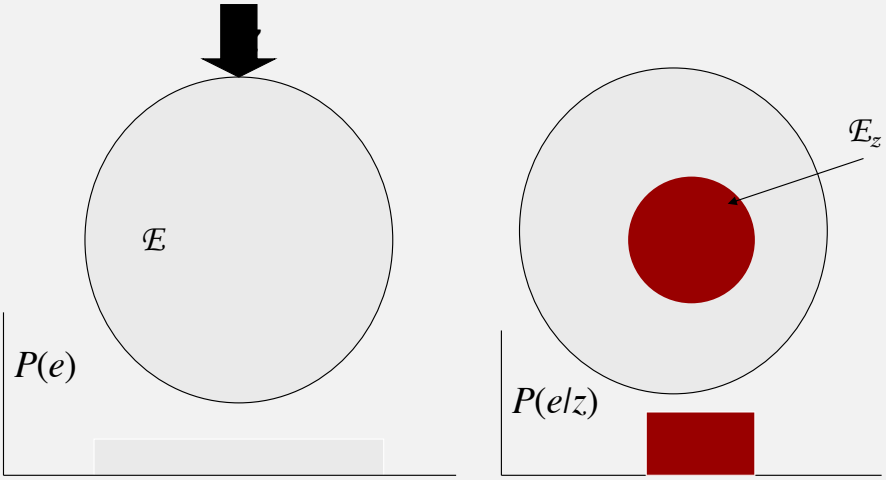
- A given piece of evidence e rules out all possible worlds that are incompatible with e or selects the possible worlds in which e is $True$. Evidence e *induces* a new measure μ_e.

$$\mu_e(\omega) = \begin{cases} \frac{1}{P(e)}\mu(\omega) \text{ if } \omega \models e \\ 0 \text{ if } \omega \not\models e \end{cases}$$

$$P(h|e) = \sum_{\omega \models h} \mu_e(\omega) = \frac{1}{P(e)} \sum_{\omega \models h \wedge e} \mu(\omega) = \frac{P(h \wedge e)}{P(e)}$$

Effect of Evidence on Possible worlds

Evidence z e.g., (color = red) rules out some assignments of values to some of the random variables



Probability as a measure over possible worlds

- Suppose I have two coins – one a normal fair coin, and the other with 2 heads.
- I pick a coin at *random* and toss it.
- Suppose the outcome is a Head.
- What is the probability that the coin that was tossed is Fair?

$$\Omega = \{(Fair, H), (Fair, T), (Rigged, H), (Rigged, T)\}$$

$$\mu = \left\{\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0\right\}$$

$$\Pr(H) = \sum_{\omega \models H} \mu(\omega) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4}$$

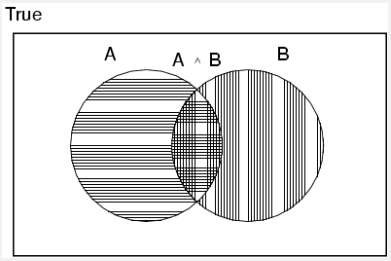
$$Pr(Fair|H) = \frac{Pr(H \wedge Fair)}{Pr(H)} = \frac{(1/4)}{(3/4)} = \frac{1}{3}$$

A note about notation

- When talking about a particular random variable value, you should technically write $P(X = x)$, etc.
- We often use the shorthand $P(x)$ to mean probability that X takes any particular value, i.e., $P(X = x)$

Properties of probabilities

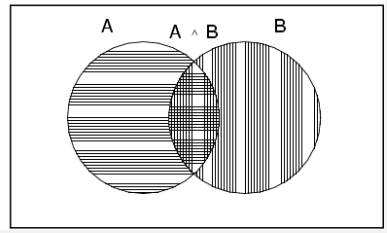
P(*A or B*) = ?



Properties of probabilities

$$P(A \text{ or } B) = P(A) + P(B) - P(A, B)$$

True



# of probabilities

$$P(\neg E) = 1 - P(E)$$

y:

ents $E = e_1, e_2, \ldots, e_n$

$$p(e_i) = 1 - \sum_{j=1:n, j \neq i} p(e_j)$$

$$P(E1, E2) \leq P(E1)$$

Chain rule (aka product rule)

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \quad \Rightarrow \quad P(X,Y) = P(X|Y)P(Y)$$

We can view calculating the probability of X *AND* Y occurring in two steps:

1. Y occurs with some probability P(Y)
2. Then, X occurs, given that Y has occurred

Chain rule

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

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

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

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

$$P(X_1,X_2 \cdots X_n) = \prod_{i=1}^{n} P(X_i|X_1 \cdots X_{i-1})$$

Applications of the chain rule

We saw that we could calculate the individual prior probabilities using the joint distribution

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

What if we don't have the joint distribution, but do have conditional probability information:

- $P(Y)$
- $P(X|Y)$

$$p(x) = \sum_{y \in Y} p(y)p(x \mid y)$$

This is called "summing over" or "marginalizing out" a variable

Bayes' rule (theorem)

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \Rightarrow P(X,Y) = P(X|Y)P(Y)$$

$$P(Y|X) = \frac{P(X,Y)}{P(X)} \Rightarrow P(X,Y) = P(Y|X)P(X)$$

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

Bayes' rule

- $p(disease \mid symptoms)$
  - For everyone who had the symptoms, what fraction had the disease?
- $p(symptoms|disease)$
  - For everyone that had the disease, what fraction had this symptom?
- $p(\ label|\ features\ )$
  - For all examples that had those features, what fraction had that label?
- $p(features \mid label)$
  - For all the examples with that label, what fraction had this feature

Bayes Rule

Does patient have cancer or not?

- A patient takes a lab test and the result comes back positive.
- The test returns
  - a correct positive result in only 98% of the cases in which the disease is actually present, and
  - a correct negative result in only 97% of the cases in which the disease is not present.
  - Furthermore, .008 fraction of the entire population have this cancer.

$$P(cancer) = \qquad P(\neg cancer) =$$
$$P(+ \mid cancer) = \qquad P(- \mid cancer) =$$
$$P(+ \mid \neg cancer) = \qquad P(- \mid \neg cancer) =$$

Bayes Rule

Does patient have cancer or not?

$$P(cancer) = 0.008 \qquad P(\neg cancer) = 0.992$$

$$P(+ \mid cancer) = 0.98 \qquad P(- \mid cancer) = 0.02$$

$$P(+ \mid \neg cancer) = 0.03 \qquad P(- \mid \neg cancer) = 0.97$$

$$P(cancer|+) = \frac{P(+|cancer)P(cancer)}{P(+)};$$

$$P(\neg cancer|+) = \frac{P(+|\neg cancer)P(\neg cancer)}{P(+)}$$

$$P(cancer|+)P(+) = 0.98 \times 0.008 = 0.0078;$$

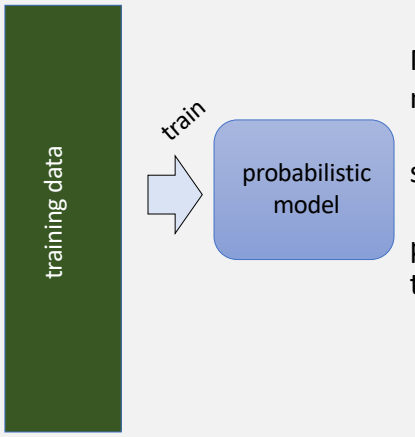
$$P(\neg cancer|+)P(+) = 0.03 \times 0.992 = 0.0298$$

$$P(+) = 0.0078 + 0.0298$$

$$P(cancer \mid +) = 0.21; \qquad P(\neg cancer \mid +) = 0.79$$

The patient, more likely than not, does not have cancer

Probabilistic Modeling



Model the data with a probabilistic model

specifically, learn p(*features, label*)

p(*features, label*) tells us how likely these features and this example are

An example: classifying fruit

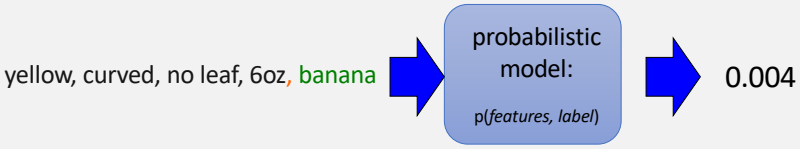
Training data

examples	label
red, round, leaf, 3oz, …	apple
green, round, no leaf, 4oz, …	apple
yellow, curved, no leaf, 4oz, …	banana
green, curved, no leaf, 5oz, …	banana

probabilistic model:
$p(features, label)$

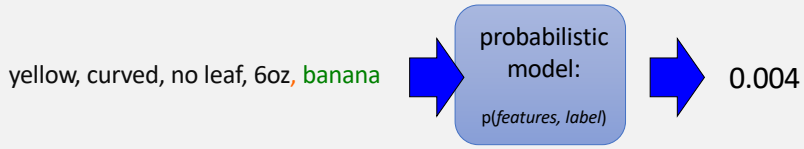
Probabilistic models

Probabilistic models define a probability distribution over features and labels:

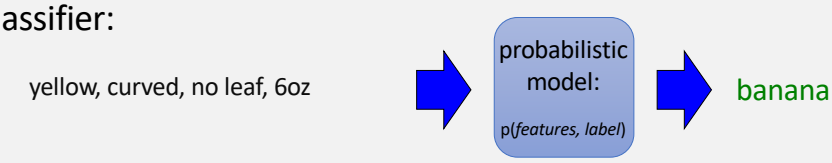


Probabilistic model vs. classifier

Probabilistic model:

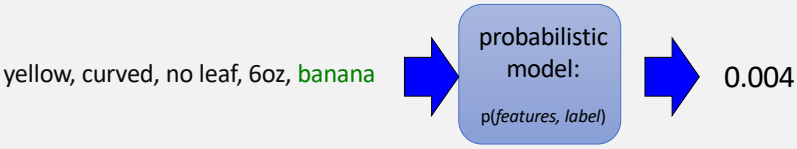


Classifier:



Probabilistic models: classification

Probabilistic models specify a probability distribution over features and labels:

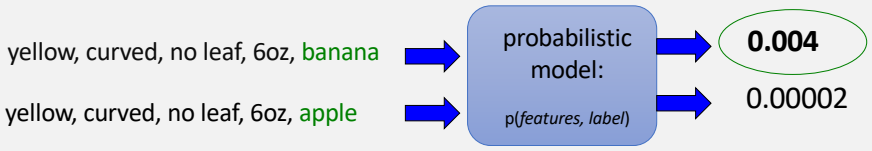


Given an unlabeled example: yellow, curved, no leaf, 6oz predict the label

How do we use a probabilistic model for classification/prediction?

Probabilistic models

Probabilistic models define a probability distribution over features and labels:



- For each label, ask for the probability under the model
- Pick the label with the highest probability

Probabilistic model vs. classifier

Probabilistic model:

yellow, curved, no leaf, 6oz, banana → probabilistic model: p(*features, label*) → 0.004

Classifier:

yellow, curved, no leaf, 6oz → probabilistic model: p(*features, label*) → banana

Why probabilistic models?

Probabilistic models

- Naturally model uncertainty
- Can be combined in well-understood ways
- Rest on strong mathematical foundations
- Can be used to generate data

Probabilistic models: questions

- Which model do we use, i.e. how do we calculate p(*feature, label*)?
- How do train the model, i.e. how to we we estimate the probabilities for the model?
- How do we deal with over-fitting?

Same problems we've been dealing with so far

Probabilistic models

- Which model do we use, i.e. how do we obtain $p(feature, label)$?
- How do train the model, i.e. how to we we estimate the probabilities for the model?
- How do we deal with over-fitting?

ML in general

- Which model do we use (linear model, non-parametric)
- How do train the model?
- How do we deal with over-fitting?

Basic steps for probabilistic modeling

Step 1: pick a model

Step 2: figure out how to estimate the probabilities for the model

Step 3: deal with overfitting

Probabilistic models

- Which model do we use, i.e. how do we calculate $p(feature, label)$?
- How do train the model, i.e. how to we we estimate the probabilities for the model?
- How do we deal with over-fitting?

Basic steps for probabilistic modeling

Step 1: pick a model

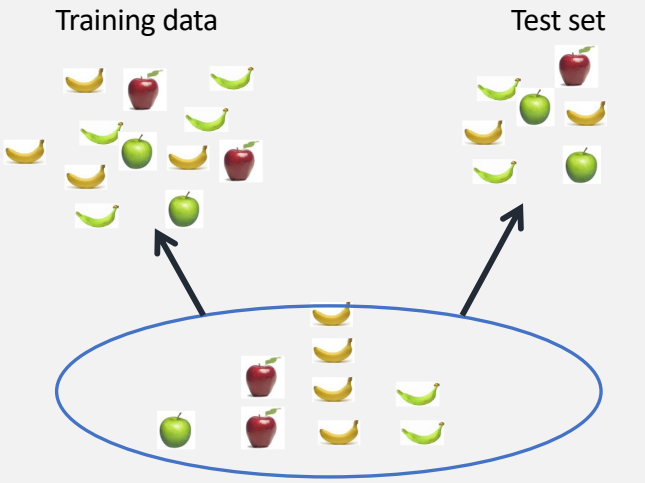
Step 2: figure out how to estimate the probabilities for the model

Step 3 (optional): deal with overfitting

Probabilistic models

- Which model do we use, i.e. how do we calculate $p(feature, label)$?
- How do train the model, i.e. how to we we estimate the probabilities for the model?
- How do we deal with overfitting?

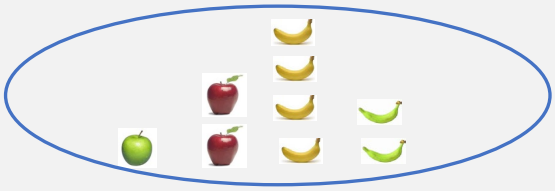
What was the data generating distribution?



Mother nature: data generating distribution

Step 1: picking a model

What we're really trying to do is model the data generating distribution – the joint distribution of *features* and labels



Mother nature: data generating distribution

Some math

$$p(features, label) = p(x_1, x_2, \cdots x_n, y)$$
$$= p(y)\, p(x_1, x_2, \cdots x_n |\, y)$$
$$= p(y) p(x_1|y) p(x_2, \cdots x_n |\, y)$$
$$= p(y) p(x_1|y) p(x_{2|}|y, x_1) p(x_3 \cdots x_n |\, y)$$
$$= p(y) \prod_{i=1}^{n} p(\, x_i |\, y, x_1 \cdots x_{i-1})$$

- chain rule!

Step 1: pick a model

$$p(features, label) = p(x_1, x_2, \cdots x_n, y)$$
$$= p(y) \prod_{i=1}^{n} p(x_i | y, x_1 \cdots x_{i-1})$$

So, far we have made NO assumptions about the data

$$p(x_m | y, x_1, x_2, \ldots, x_{m-1})$$

- How many entries would the probability distribution table have?
- Suppose we have 10000 binary features and a binary label?

2^{10001}

Decision Theoretic Foundations of an Optimal Probabilistic Classifier

- What is an "optimal" classifier?
- How can a classifier assign labels optimally?
- Can we build an optimal classifier?
- Example of an optimal classifier

Decision theoretic foundations of classification

Consider the problem of classifying an instance X
into one of two mutually exclusive classes ω_1 or ω_2

$P(\omega_1 | X)$ = probability of class ω_1 given the evidence X

$P(\omega_2 | X)$ = probability of class ω_2 given the evidence X

What is the probability of error?

$P(error | X) = P(\omega_1 | X)$ if we choose ω_2

$= P(\omega_2 | X)$ if we choose ω_1

Minimum Error Classification

To minimize classification error

Choose ω_1 if $P(\omega_1|X) > P(\omega_2|X)$

Choose ω_2 if $P(\omega_2|X) > P(\omega_1|X)$

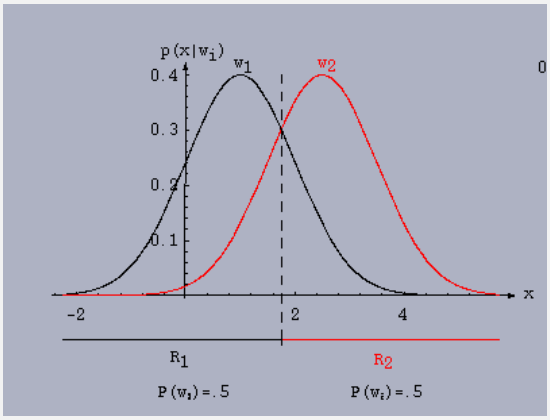
which yields

$$P(error \mid X) = \min\left[P(\omega_1|X), P(\omega_2|X)\right]$$

We have :

$$P(\omega_1|X) = P(X \mid \omega_1)P(\omega_1);$$

$$P(\omega_2|X) = P(X \mid \omega_2)P(\omega_2)$$



Choose ω_1 if $P(\omega_1|X) > P(\omega_2|X)$ i.e. $X \in R_1$

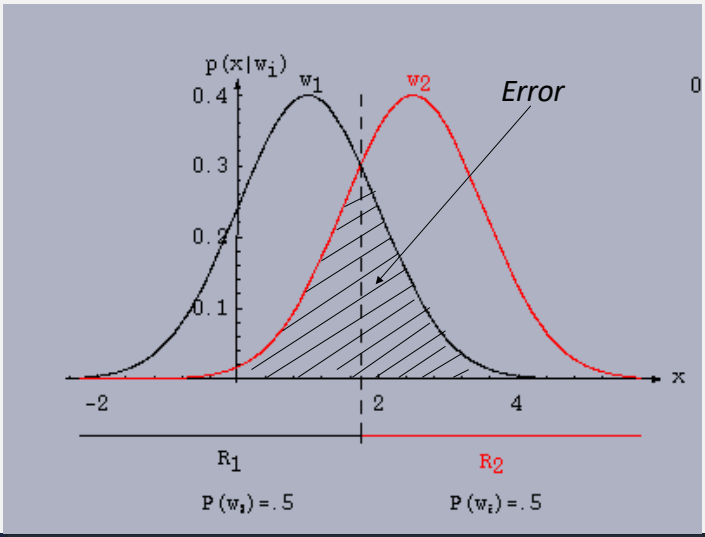
Choose ω_2 if $P(\omega_2|X) > P(\omega_1|X)$ i.e. $X \in R_2$

Optimality of Bayes Decision Rule

We can show that the Bayesian classifier

- is optimal in that it is guaranteed to minimize the probability of misclassification

Optimality of Bayes Decision Rule



Optimality of Bayes Decision Rule

- The result generalizes to multivariate input spaces
- Similar result can be proved in the case of discrete (as opposed to continuous) input spaces – replace integration over the input space by summation

Bayes Decision Rule yields Minimum Error Classification

To minimize classification error

Choose ω_1 if $P(\omega_1|X) > P(\omega_2|X)$

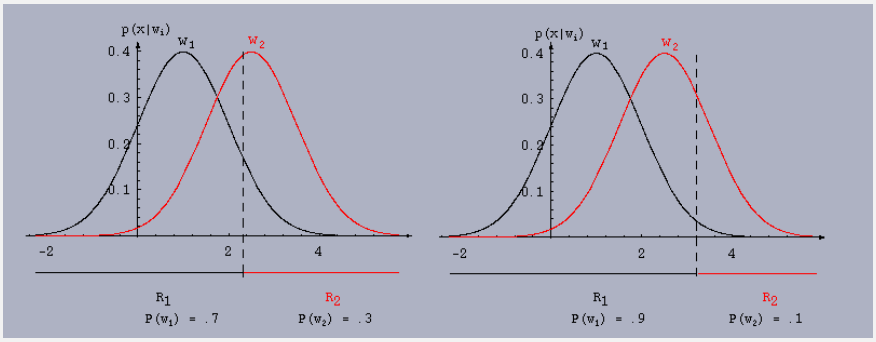
Choose ω_2 if $P(\omega_2|X) > P(\omega_1|X)$

which yields

$$P(error \mid X) = \min\left[P(\omega_1|X), P(\omega_2|X)\right]$$

Bayes Decision Rule

Behavior of Bayes decision rule as a function of prior probability of classes



Bayes Optimal Classifier

Classification rule that guarantees minimum error :

Choose ω_1 if $P(X \mid \omega_1)P(\omega_1) > P(X \mid \omega_2)P(\omega_2)$

Choose ω_2 if $P(X \mid \omega_2)P(\omega_2) > P(X \mid \omega_1)P(\omega_1)$

If $P(X \mid \omega_1) = P(X \mid \omega_2)$

classification depends entirely on $P(\omega_1)$ and $P(\omega_2)$

If $P(\omega_1) = P(\omega_2)$,

classification depends entirely on $P(X \mid \omega_1)$ and $P(X \mid \omega_2)$

Bayes classification rule combines the effect of the two terms optimally - so as to yield minimum error classification.

Generalization to multiple classes $c(X) = \arg\max_{\omega_j} P(\omega_j \mid X)$

Minimum Risk Classification

Let λ_{ij} = risk or cost associated with assigning an instance to class ω_j when the correct classification is ω_i

$R(\omega_i \mid X)$ = expected loss incurred in assigning X to class ω_i

$$R(\omega_1 \mid X) = \lambda_{11} P(\omega_1 \mid X) + \lambda_{21} P(\omega_2 \mid X)$$

$$R(\omega_2 \mid X) = \lambda_{12} P(\omega_1 \mid X) + \lambda_{22} P(\omega_2 \mid X)$$

Classification rule that guarantees minimum risk:

Choose ω_1 if $R(\omega_1 \mid X) < R(\omega_2 \mid X)$

Choose ω_2 if $R(\omega_2 \mid X) < R(\omega_1 \mid X)$

Flip a coin otherwise

Minimum Risk Classification

λ_{ij} = risk or cost associated with assigning an instance to class ω_j when the correct classification is ω_i

Ordinarily $(\lambda_{21} - \lambda_{22})$ and $(\lambda_{12} - \lambda_{11})$ are positive

(cost of being correct is less than the cost of error)

So we choose ω_1 if $\frac{P(X|\omega_1)}{P(X|\omega_2)} > \frac{(\lambda_{21} - \lambda_{22})}{(\lambda_{12} - \lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)}$

Otherwise choose ω_2

Minimum error classification rule is a special case :

$$\lambda_{ij} = 0 \text{ if } i = j \text{ and } \lambda_{ij} = 1 \text{ if } i \neq j$$

This classification rule can be shown to be optimal in that it is guaranteed to minimize the risk of misclassification

Summary of Bayesian recipe for classification

λ_{ij} = risk or cost associated with assigning an instance to class ω_j when the correct classification is ω_i

Choose ω_1 if $\frac{P(X|\omega_1)}{P(X|\omega_2)} > \frac{(\lambda_{21}-\lambda_{22})}{(\lambda_{12}-\lambda_{11})}\frac{P(\omega_2)}{P(\omega_1)}$

Choose ω_2 if $\frac{P(X|\omega_1)}{P(X|\omega_2)} < \frac{(\lambda_{21}-\lambda_{22})}{(\lambda_{12}-\lambda_{11})}\frac{P(\omega_2)}{P(\omega_1)}$

Minimum error classification rule is a special case :

Choose ω_1 if $\frac{P(X|\omega_1)}{P(X|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$ Otherwise choose ω_2

Bayesian recipe for classification

Note that $P(\omega_i \mid \mathbf{x}) = \dfrac{P(\mathbf{x}|\omega_i)P(\omega_i)}{P(\mathbf{x})}$

Model $P(\mathbf{x} \mid \omega_1)$, $P(\mathbf{x}|\omega_2)$, $P(\omega_1)$, and $P(\omega_2)$

Using Bayes rule, choose ω_1 if $P(\mathbf{x} \mid \omega_1)P(\omega_1) > P(\mathbf{x}|\omega_2)P(\omega_2)$

Otherwise choose ω_2

Multiple disjoint classes

Estimate $P(\omega_i|X) = \frac{P(X|\ \omega_i)P(\omega_i)}{P(X)}$

$\omega = argmax\ P(\omega_i|X)$

Assign sample to the most probable class!

Summary of Bayesian recipe for classification

- The Bayesian recipe is simple, optimal, and in principle, straightforward to apply
- To use this recipe in practice, we need to know $P(X|\omega_i)$ – the generative model for data for each class and $P(\omega_i)$ – the prior probabilities of classes
- Because these probabilities are unknown, we need to estimate them from data – or learn them!
- X is typically high-dimensional or may have complex structure
- Need to estimate $P(X|\omega_i)$ from data