浙江大学伊利诺伊大学厄巴纳香槟校区联合学院

Zhejiang University-University of Illinois at Urbana Champaign Institute

ECE448: Artificial Intelligence

Lecture 13: Bayesian Inference and Bayesian Learning

Prof. Hongwei Wang hongweiwang@intl.zju.edu.cn

Prof. Mark Hasegawa-Johnson jhasegaw@illinois.edu

Spring 2023

Outline：Bayesian Inference and Bayesian Learning

ZJUI

1. **Bayes Rule**
2. **Bayesian Inference**
   - **Misdiagnosis**
   - **The Bayesian “Decision”**
   - **The “Naïve Bayesian” Assumption**
   - **Bag of Words (BoW)**
3. **Bayesian Learning**
   - **Maximum Likelihood estimation of parameters**
   - **Maximum A Posteriori estimation of parameters**
   - **Laplace Smoothing**

Outline：Bayesian Inference and Bayesian Learning

ZJUI

1. **Bayes Rule**
2. **Bayesian Inference**
   - **Misdiagnosis**
   - **The Bayesian “Decision”**
   - **The “Naïve Bayesian” Assumption**
   - **Bag of Words (BoW)**
3. **Bayesian Learning**
   - **Maximum Likelihood estimation of parameters**
   - **Maximum A Posteriori estimation of parameters**
   - **Laplace Smoothing**

Bayes' Rule

ZJUI

Rev. Thomas Bayes
(1702-1761)

- The product rule gives us two ways to factor a joint probability:

$$P(A,B) = P(B|A)P(A) = P(A|B)P(B)$$

- Therefore,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Why is this useful?
  - "A" is something we care about, but P(A|B) is really really hard to measure (example: the sun exploded)
  - "B" is something less interesting, but P(B|A) is easy to measure (example: the amount of light falling on a solar cell)
  - Bayes' rule tells us how to compute the probability we want (P(A|B)) from probabilities that are much, much easier to measure (P(B|A)).

Bayes Rule example

Eliot & Karson are getting married tomorrow, at an outdoor ceremony in the desert.

- In recent years, it has rained only 5 days each year (5/365 = 0.014).
$$P(R) = 0.014$$
- Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time.
$$P(F|R) = 0.9$$
- When it doesn't rain, he incorrectly forecasts rain 10% of the time.
$$P(F|\neg R) = 0.1$$
- What is the probability that it will rain on Eliot's wedding?

$$P(R|F) = \frac{P(F|R)P(R)}{P(F)} = \frac{P(F,R)P(R)}{P(F,R) + P(F,\neg R)} = \frac{P(F|R)P(R)}{P(F|R)P(R) + P(F|\neg R)P(\neg R)}$$

$$= \frac{(0.9)(0.014)}{(0.9)(0.014) + (0.1)(0.956)} = 0.116$$

The More Useful Version of Bayes' Rule

Rev. Thomas Bayes
(1702-1761)

This version is what you memorize.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- Remember, P(B|A) is easy to measure (the probability that light hits our solar cell, if the sun still exists and it's daytime). Let's assume we also know P(A) (the probability the sun still exists).
- But suppose we don't really know P(B) (what is the probability light hits our solar cell, if we don't really know whether the sun still exists or not?)

This version is what you actually use.

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

ZJUI

1. Bayes Rule
2. Bayesian Inference
    - Misdiagnosis
    - The Bayesian "Decision"
    - The "Naïve Bayesian" Assumption
    - Bag of Words (BoW)
3. Bayesian Learning
    - Maximum Likelihood estimation of parameters
    - Maximum A Posteriori estimation of parameters
    - Laplace Smoothing

The Misdiagnosis Problem

1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammographies. 9.6% of women without breast cancer will also get positive mammographies. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

$$P(\text{cancer} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive})}$$

$$= \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive} \mid \text{cancer})P(\text{cancer}) + P(\text{positive} \mid \neg\text{cancer})P(\neg\text{Cancer})}$$

$$= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.096 \times 0.99} = \frac{0.008}{0.008 + 0.095} = 0.0776$$

Second Opinions

If your doctor tells you that you have a health problem or suggests a treatment for an illness or injury, you might want a second opinion. This is especially true when you're considering surgery or major procedures.

Asking another doctor to review your case can be useful for many reasons:

- Doctors have different styles. Some may be more likely to suggest surgery or other major treatments. Others may suggest a slower, wait-and-see approach. Getting a second opinion can help you weigh the pros and cons of their treatment plans.
- You can be well-informed before you make a health decision. Another opinion allows you to discuss your options with a qualified doctor. For example, you may have to choose between traditional or robotic surgery. It's good to think about the benefits and risks of both types. Or you might be considering different types of cancer treatment and want to visit several hospitals. Or another doctor's opinion might shed more light on your diagnosis. The extra opinions help you make educated

HEALTH INSURANCE AND MEDICARE HOME

News
Reference
Quizzes
Videos
Message Boards
Find a Doctor

RELATED TO HEALTH INSURANCE AND MEDICARE

Health Insurance Terms
Insurance Myths and Facts
Using Your Benefits
Copay vs. Coinsurance
Screening Tests
Getting a Second Opinion
FSA vs. HSA
Nursing Home Care
Help Paying for Rx

TODAY ON WEBMD


Clinical Trials
What qualifies you for one?


Working During Cancer Treatment
Know your benefits.


Going to the Dentist?
How to save money.


Enrolling in Medicare
How to get started.

The Bayesian Decision

The agent is given some evidence, E.

The agent has to make a decision about the value of an unobserved variable Y. Y is called the “query variable” or the “class variable” or the “category.”

- Partially observable, stochastic, episodic environment
- Example: $Y \in$ {spam, not spam}, E = email message.
- Example: $Y \in$ {zebra, giraffe, hippo}, E = image features

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

- The query variable, Y, is a random variable. Assume its pmf, P(Y=y) is known.
- Furthermore, the true value of Y has already been determined --- we just don't know what it is!
- The agent must ACT by saying "I believe that Y=a".
- The agent has a **post-hoc loss function** $L(y,a)$
  - $L(y,a)$ is the loss if the true value is Y=y, but the agent says "a"
- The **a priori loss function** $L(Y,a)$ is a binary random variable
  - $P(L(Y,a)=0) \ = \ P(Y=a)$
  - $P(L(Y,a)=1) \ = \ P(Y \neq a)$

Loss Function Example

- Suppose Y=outcome of a coin toss.
- The agent will choose the action "a" (which is either a=heads, or a=tails)
- The loss function L(y,a) is

L(y,a)	y=heads	y=tails
a=heads	0	1
a=tails	1	0

- Suppose we know that the coin is biased, so that P(Y=heads)=0.6. Therefore the agent chooses a=heads. The loss function L(Y,a) is now a random variable:

	c=0	c=1
P(L(Y,a)=c)	0.6	0.4

- The observation, E, is another random variable. Suppose the joint probability $P(Y = y, E = e)$ is known.
- The agent is allowed to observe the true value of E=e before it guesses the value of Y.
- Suppose that the observed value of E is E=e. Suppose the agent guesses that Y=a. Then its loss, L(Y,a), is a conditional random variable:

$$P(L(Y,a) = 0|E = e) = P(Y = a|E = e)$$

$$P(L(Y,a) = 1|E = e) = P(Y \neq a|E = e) = \sum_{y \neq a} P(Y = y|E = e)$$

- Suppose the agent chooses any particular action "a", then its expected loss is:

$$E[L(Y,a)|E=e] = \sum_{y} L(y,a)P(Y=y|E=e) = \sum_{y\neq a} P(Y=y|E=e)$$

- Which action, "a", should the agent choose in order to minimize its expected loss?
- The one that has the greatest posterior probability. The best value of "a" to choose is the one given by:

$$a = \arg\max_{a} P(Y=a|E=e)$$

- This is called the Maximum a Posteriori (MAP) decision

MAP decision

The action, “a”, should be the value of C that has the highest posterior probability given the observation X=x:

$$a = \operatorname{argmax} P(Y = a|E = e) = \operatorname{argmax} \frac{P(E = e|Y = a)P(Y = a)}{P(E = e)}$$
$$= \operatorname{argmax} P(E = e|Y = a)P(Y = a)$$

$$\underbrace{P(Y = a|E = e)}_{\text{posterior}} \propto \underbrace{P(E = e|Y = a)}_{\text{likelihood}}\underbrace{P(Y = a)}_{\text{prior}}$$

- Maximum Likelihood (ML) decision:

$$a = \operatorname{argmax} P(E = e|Y = a)$$

The Bayesian Terms

- $P(Y = y)$ is called the "**prior**" (*a priori*, in Latin) because it represents your belief about the query variable before you see any observation.
- $P(Y = y|E = e)$ is called the "**posterior**" (*a posteriori*, in Latin), because it represents your belief about the query variable after you see the observation.
- $P(E = e|Y = y)$ is called the "**likelihood**" because it tells you how much the observation, E=e, is like the observations you expect if Y=y.
- $P(E = e)$ is called the "**evidence distribution**" because E is the evidence variable, and $P(E = e)$ is its marginal distribution.

$$P(y|e) = \frac{P(e|y)P(y)}{P(e)}$$

1. **Bayes Rule**
2. **Bayesian Inference**
   - **Misdiagnosis**
   - **The Bayesian “Decision”**
   - **The “Naïve Bayesian” Assumption**
   - **Bag of Words (BoW)**
3. **Bayesian Learning**
   - **Maximum Likelihood estimation of parameters**
   - **Maximum A Posteriori estimation of parameters**
   - **Laplace Smoothing**

Naïve Bayes model

- Suppose we have many different types of observations (symptoms, features) $X_1, \ldots, X_n$ that we want to use to obtain evidence about an underlying hypothesis C
- MAP decision:

$$P(Y = y|E_1 = e_1, \ldots, E_n = e_n) \propto$$
$$P(Y = y)P(E_1 = e_1, \ldots, E_n = e_n|Y = y)$$

- If each feature E_i can take on k values, how many entries are in the pmf table $P(E_1 = e_1, \ldots, E_n = e_n|Y = y)$?

Naïve Bayes model

- How many entries are in the pmf table $P(e_1, \dots, e_n|y)$?
  - Without naïve Bayes: $k(k^n - 1)$
  - (k values of $\mathrm{Y} = y, \quad k(k^n - 1)$ possible combinations of $e_1, \dots, e_n$)
- We can make the simplifying assumption that the different features are conditionally independent *given the hypothesis*:

$$P(e_1, \dots, e_n|y) \approx P(e_1|y)P(e_2|y) \dots P(e_n|y)$$

- If each observation and the hypothesis can take on *k* values, what is the complexity of storing the resulting distributions?
  - Each $P(e_i|y)$ requires $(k-1)\times k \quad$ (k values of $\mathrm{Y} = y$, $k-1$ of $E_i = e_i$)
  - There are n of them, for a total space requirement: $n\times(k-1)\times k$

Naïve Bayes model

Suppose we have many different types of observations (symptoms, features) $E_1, \ldots, E_n$ that we want to use to obtain evidence about an underlying hypothesis Y

MAP decision:

$$a = \operatorname{argmax} p(Y = a | E_1 = e_1, \ldots, E_n = e_n)$$

$$= \operatorname{argmax} p(Y = a) p(E_1 = e_1, \ldots, E_n = e_n | Y = a)$$

$$\approx \operatorname{argmax} p(Y = a) p(y_1 | a) p(y_2 | a) \ldots p(y_n | a)$$

Case study: Text document classification

- **MAP decision:** assign a document to the class with the highest posterior P(class | document)

- Example: spam classification
  - Classify a message as spam if P(spam | message) > P(¬spam | message)

> Dear Sir.
>
> First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

> TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.
>
> 99 MILLION EMAIL ADDRESSES FOR ONLY $99

> Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Case study: Text document classification

- **MAP decision:** assign a document to the class with the highest posterior P(class | document)

- We have P(class | document) $\propto$ P(document | class)P(class)

- To enable classification, we need to be able to estimate the likelihoods P(document | class) for all classes and
priors P(class)

Naïve Bayes Representation

- Goal: estimate likelihoods P(document | class) and priors P(class)
- Likelihood: ***bag of words*** representation
  - The document is a sequence of words $(w_1, \ldots, w_n)$
  - The order of the words in the document is not important
  - Each word is conditionally independent of the others given document class

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Naïve Bayes Representation

- Goal: estimate likelihoods P(document | class) and priors P(class)
- Likelihood: ***bag of words*** representation
  - The document is a sequence of words $(E_1 = w_1, \ldots, E_n = w_n)$
  - The order of the words in the document is not important
  - Each word is conditionally independent of the others given document class

$$P(document \mid class) = P(w_1, \ \ldots \ , w_n \mid class) = \prod_{i=1}^{n} P(w_i \mid class)$$

  - Thus, the problem is reduced to estimating marginal likelihoods of individual words $p(w_i \mid class)$

Parameter estimation

- Model parameters: feature likelihoods p(word | class) and priors p(class)
  - How do we obtain the values of these parameters?

prior

spam:	0.33
¬spam:	0.67

P(word | spam)

the :	0.0156
to :	0.0153
and :	0.0115
of :	0.0095
you :	0.0093
a :	0.0086
with:	0.0080
from:	0.0075
...	

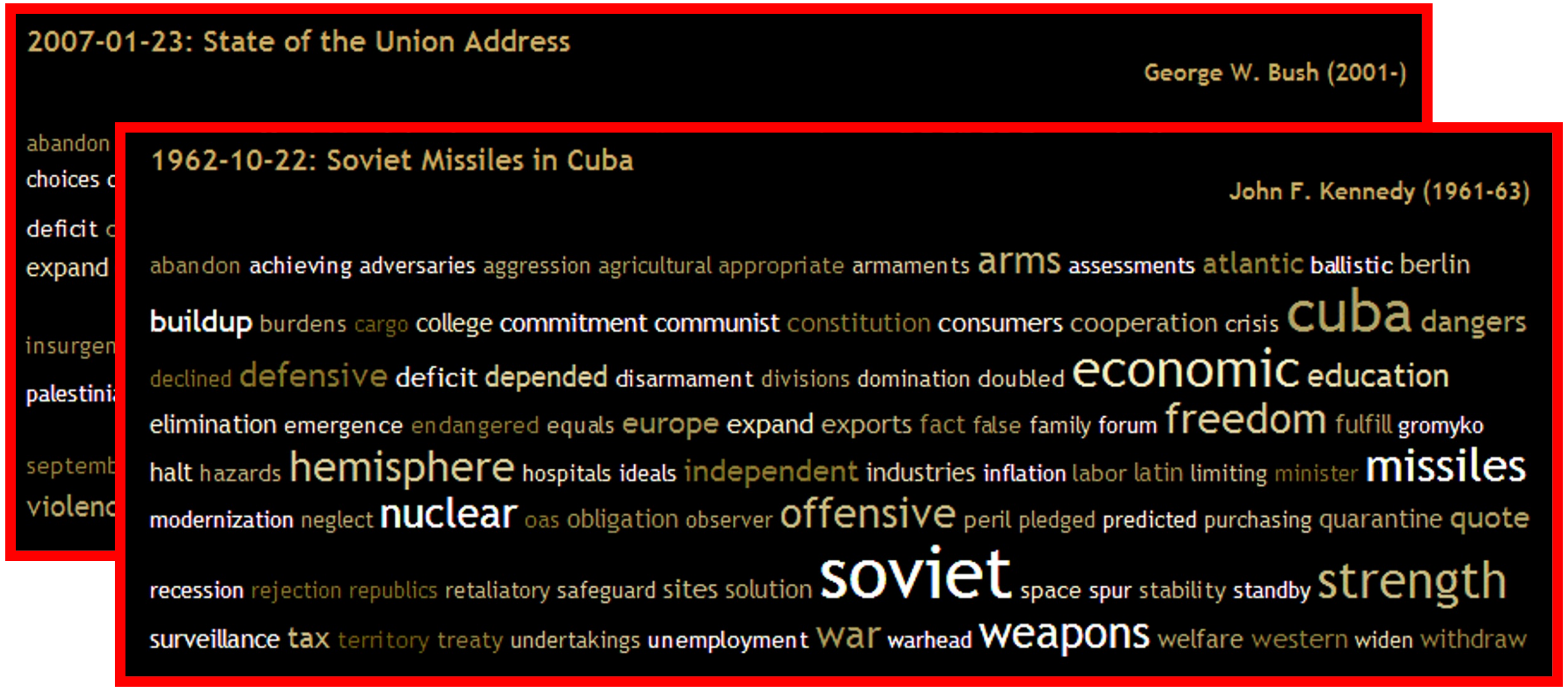
P(word | ¬spam)

the :	0.0210
to :	0.0133
of :	0.0119
2002:	0.0110
with:	0.0108
from:	0.0107
and :	0.0105
a :	0.0100
...	

Bag of words illustration

US Presidential Speeches Tag Cloud
http://chir.ag/projects/preztags/

Bag of words illustration



US Presidential Speeches Tag Cloud
http://chir.ag/projects/preztags/

Bag of words illustration

US Presidential Speeches Tag Cloud
http://chir.ag/projects/preztags/

ZJUI

1. Bayes Rule
2. Bayesian Inference
   - Misdiagnosis
   - The Bayesian "Decision"
   - The "Naïve Bayesian" Assumption
   - Bag of Words (BoW)
3. Bayesian Learning
   - Maximum Likelihood estimation of parameters
   - Maximum A Posteriori estimation of parameters
   - Laplace Smoothing

- Model parameters: feature likelihoods P(word | class) and priors P(class)
  - How do we obtain the values of these parameters?
  - Need *training set* of labeled samples from both classes

$$P(\text{word} \mid \text{class}) = \frac{\text{\# of occurrences of this word in docs from this class}}{\text{total \# of words in docs from this class}}$$

- This is the *maximum likelihood* (ML) estimate, or estimate that maximizes the likelihood of the training data:

$$\prod_{d=1}^{D} \prod_{i=1}^{n_d} P(w_{d,i} \mid class_{d,i})$$

d: index of training document, *i*: index of a word

- The “bag of words model” has the following parameters:
  - $\lambda_{cw} \equiv P(W = w|C = c)$
  - $\pi_c \equiv P(C = c)$
- The training data are a set of documents, $E = [D_1, \dots, D_m]$, each with its associated class label, $Y = [C_1, \dots, C_m]$.
- The likelihood of the training data is the probability of its observations given its labels. If we assume that each document is independent of the others (“episodic”), then we get:

$$P(E, Y) = \prod_{i=1}^{m} P(D_i|C_i)P(C_i)$$

- The “bag of words model” has the following parameters:
  - $\lambda_{cw} \equiv P(W = w|C = c)$
  - $\pi_c \equiv P(C = c)$
- Each document is a sequence of words, $D_i = [W_{1i}, \dots, W_{ni}]$.
- If we assume that each word is conditionally independent given the class (the naïve Bayes a.k.a. bag-of-words assumption), then we get:

$$P(E,Y) = \prod_{i=1}^{m} P(C_i = c_i) \prod_{j=1}^{n} P(W_{ji} = w_{ji}|C_i = c_i) = \prod_{i=1}^{m} \pi_{c_i} \prod_{j=1}^{n} \lambda_{c_i w_{ji}}$$

The data likelihood $P(X, Y)$ is maximized if we choose:

$$\lambda_{cw} = \frac{\text{\# occurrences of word } w \text{ in documents of type } c}{\text{total number of words in all documents of type } c}$$

$$\pi_c = \frac{\text{\# documents of type } c}{\text{total number of documents}}$$

What is the probability that the sun will fail to rise tomorrow?

- # times we have observed the sun to rise = 100,000,000
- # times we have observed the sun not to rise = 0
- Estimated probability the sun will not rise = $\frac{0}{0+100{,}000{,}000} = 0$

Oops....

Laplace Smoothing

- The basic idea: add 1 “unobserved observation” to every possible event
- # times the sun has risen or might have ever risen = 100,000,000+1 = 100,000,001
- # times the sun has failed to rise or might have ever failed to rise = 0+1 = 1
- Estimated probability the sun will not rise = $\frac{1}{1+100{,}000{,}001} = 0.0000000099999998$

Parameter estimation

- ML (Maximum Likelihood) parameter estimate:

$$P(\text{word} \mid \text{class}) = \frac{\text{\# of occurrences of this word in docs from this class}}{\text{total \# of words in docs from this class}}$$

- Laplacian Smoothing estimate
  - How can you estimate the probability of a word you never saw in the training set? (Hint: what happens if you give it probability 0, then it actually occurs in a test document?)
  - **Laplacian smoothing:** pretend you have seen every vocabulary word one more time than you actually did

$$P(\text{word} \mid \text{class}) = \frac{\text{\# of occurrences of this word in docs from this class} + 1}{\text{total \# of words in docs from this class} + V}$$

(V: total number of unique words)

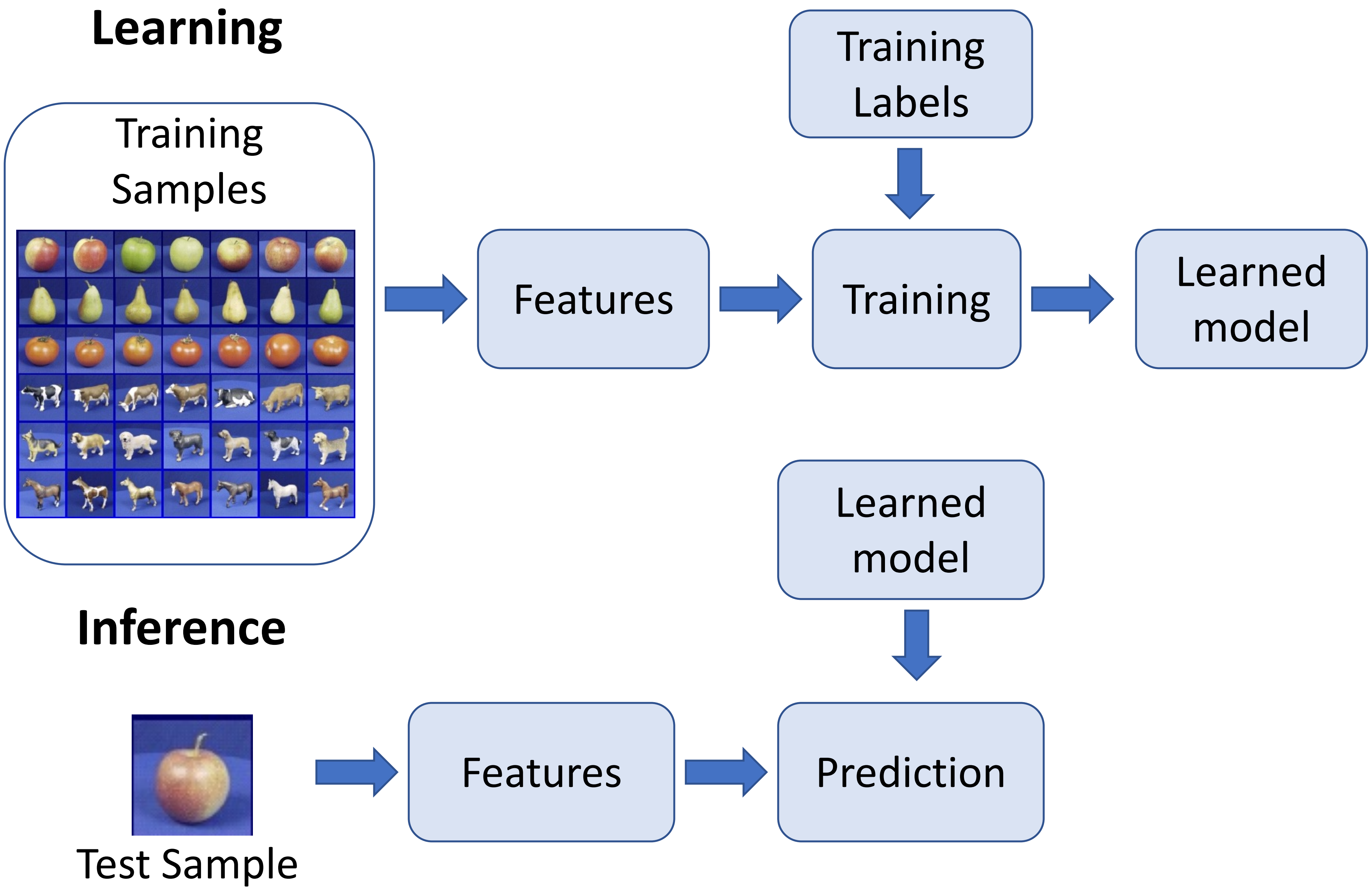
- Naïve Bayes model: assign the document to the class with the highest posterior

$$P(class \mid document) \propto P(class)\prod_{i=1}^{n} P(w_i \mid class)$$

- Model parameters:

prior	Likelihood of class 1		Likelihood of class K
$P(\text{class}_1)$	$P(w_1 \mid \text{class}_1)$		$P(w_1 \mid \text{class}_K)$
$\dots$	$P(w_2 \mid \text{class}_1)$	$\dots$	$P(w_2 \mid \text{class}_K)$
$P(\text{class}_K)$	$\dots$		$\dots$
	$P(w_n \mid \text{class}_1)$		$P(w_n \mid \text{class}_K)$

Bayesian Learning and Bayesian Inference irl:



Review: Bayesian decision making

- Suppose the agent has to make decisions about the value of an unobserved *query variable* Y based on the values of an observed *evidence variable* E
- **Inference problem:** given some observation $E = e$, what is $P(Y \mid E=e)$?
- **Learning problem:** estimate the parameters of the probabilistic model $P(y \mid e)$ given a *training sample* $\{(e_1,y_1), \ldots, (e_n,y_n)\}$