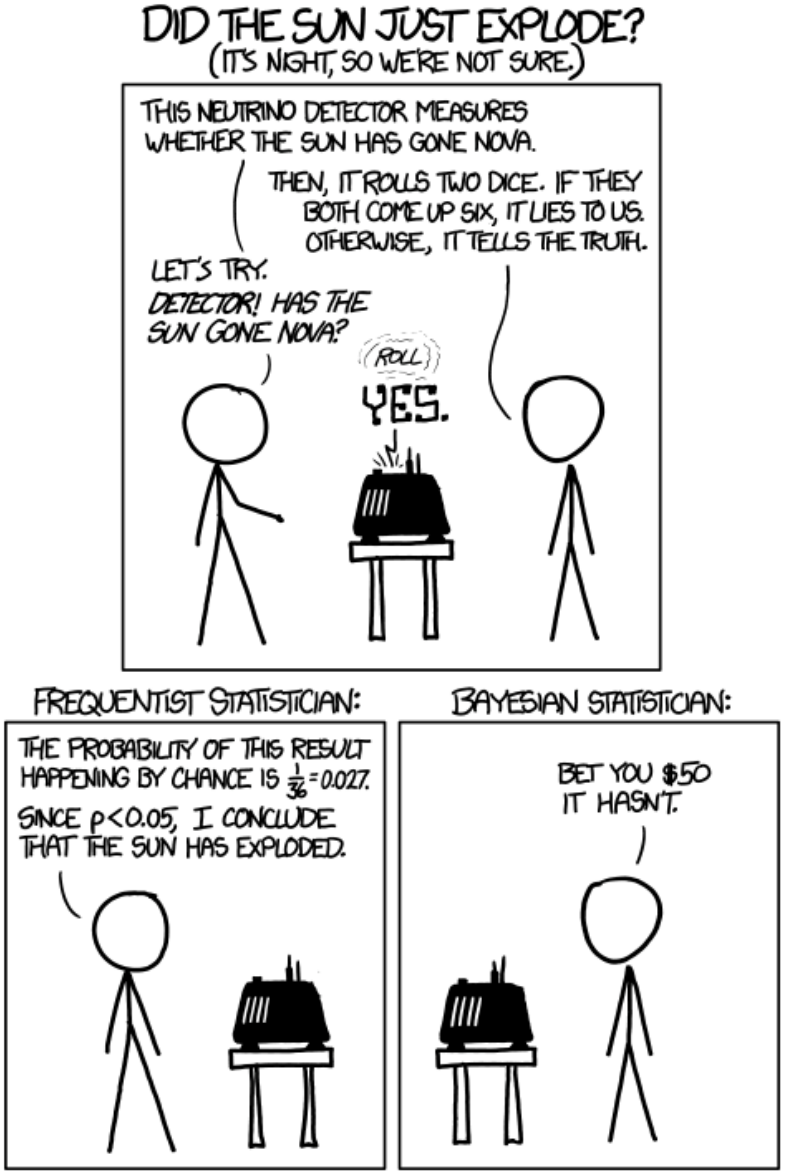
CS440/ECE448 Lecture 14: Bayesian Inference and Bayesian Learning

Slides by Svetlana Lazebnik, 10/2016

Modified by Mark Hasegawa-Johnson, 3/2018



Bayesian Inference and Bayesian Learning

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

Bayes Rule

Rev. Thomas Bayes
(1702-1761)

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

$$P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- Therefore,

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

- Why is this useful?
  - Can update our beliefs about A based on evidence B
    - P(A) is the *prior* and P(A|B) is the *posterior*
  - Key tool for probabilistic inference: can get *diagnostic probability* from *causal probability*
    - E.g., P(Cavity = true | Toothache = true) from P(Toothache = true | Cavity = true)

Bayes Rule example

Dan & Dana 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). Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on their wedding?

Bayes Rule example

Dan & Dana 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). Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on their wedding?

$$P(\text{rain} \mid \text{predict}) = \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict})}$$

$$= \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict} \mid \text{rain})P(\text{rain}) + P(\text{predict} \mid \neg\text{rain})P(\neg\text{rain})}$$

$$= \frac{0.9 \times 0.014}{0.9 \times 0.014 + 0.1 \times 0.986} = \frac{0.0126}{0.0126 + 0.0986} = 0.111$$

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian “Decision”
  - The “Naïve Bayesian” Assumption
  - Bag of Words (BoW)
- 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?

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

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

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

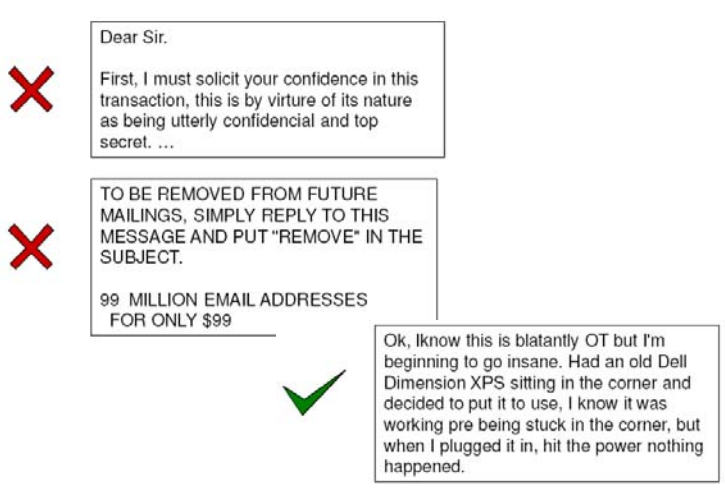
https://www.youtube.com/watch?v=BcvLAw-JRss

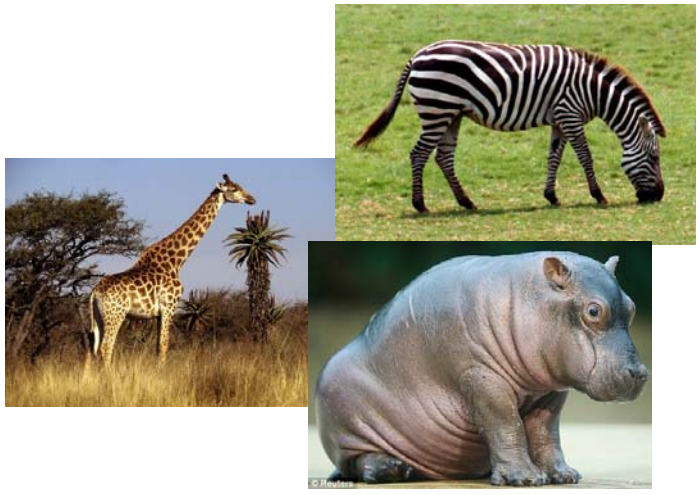
The Bayesian Decision

The agent is given some observations, x.

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

- Partially observable, stochastic, episodic environment
- Examples: C = {spam, not spam}, x = email message
  C = {zebra, giraffe, hippo}, x = image features





The Bayesian Decision

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

The Bayesian Decision

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

$$P(L(C,a) = 0|X = x) \;=\; P(C = a|X = x)$$

$$P(L(C,a) = 1|X = x) = \; P(C \neq a|X = x) = \sum_{c \neq a} P(C = c|X = x)$$

The Bayesian Decision

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

$$E[L(C,a)|X=x] = \sum_{c} L(c,a)P(C=c|X=x) = \sum_{c\neq a} P(C=c|X=x)$$

- 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:

$$a = \arg\max_{a} P(C=a|X=x)$$

- 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:

$$a = \text{argmax} P(C = a|X = x) = \text{argmax} \frac{P(X = x|C = a)P(C = a)}{P(X = x)}$$
$$= \text{argmax} P(X = x|C = a)P(C = a)$$

$$P(C = a|X = x) = P(X = x|C = a)P(C = a)$$

posterior likelihood prior

- Maximum Likelihood (ML) decision:

$$a = \text{argmax} P(X = x|C = a)$$

The Bayesian Terms

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

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian "Decision"
  - The "Naïve Bayesian" Assumption
  - Bag of Words (BoW)
- 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:

$$P(C = c | X_1 = x_1, \ldots, X_n = x_n) \propto$$
$$P(C = c) P(X_1 = x_1, \ldots, X_n = x_n | C = c)$$

- If each feature X_i can take on k values, how many entries are in the (conditional) joint probability table $P(X_1, \ldots, X_n \mid C = c)$?

Naïve Bayes model

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

$$P(C = c|X_1 = x_1, \dots, X_n = x_n) \propto$$
$$P(C = c)P(X_1 = x_1, \dots, X_n = x_n|C = c)$$

- We can make the simplifying assumption that the different features are conditionally independent *given the hypothesis*:

$$P(x_1, x_2, \dots, x_n|c) \approx P(x_1|c)P(x_2|c) \dots P(x_n|c)$$

  - If each observation and the hypothesis can take on *k* values, what is the complexity of storing the resulting distributions?
  - W.o naïve Bayes: $k(k^n - 1)$
  - With naïve Bayes: each $P(x_i|c)$ requires (k-1)*k (k values of c, k-1 of x)
  - There are n of them -> n*(k-1)*k

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:

$$a = \text{argmax} P(C = a | X_1 = x_1, \ldots, X_n = x_n)$$

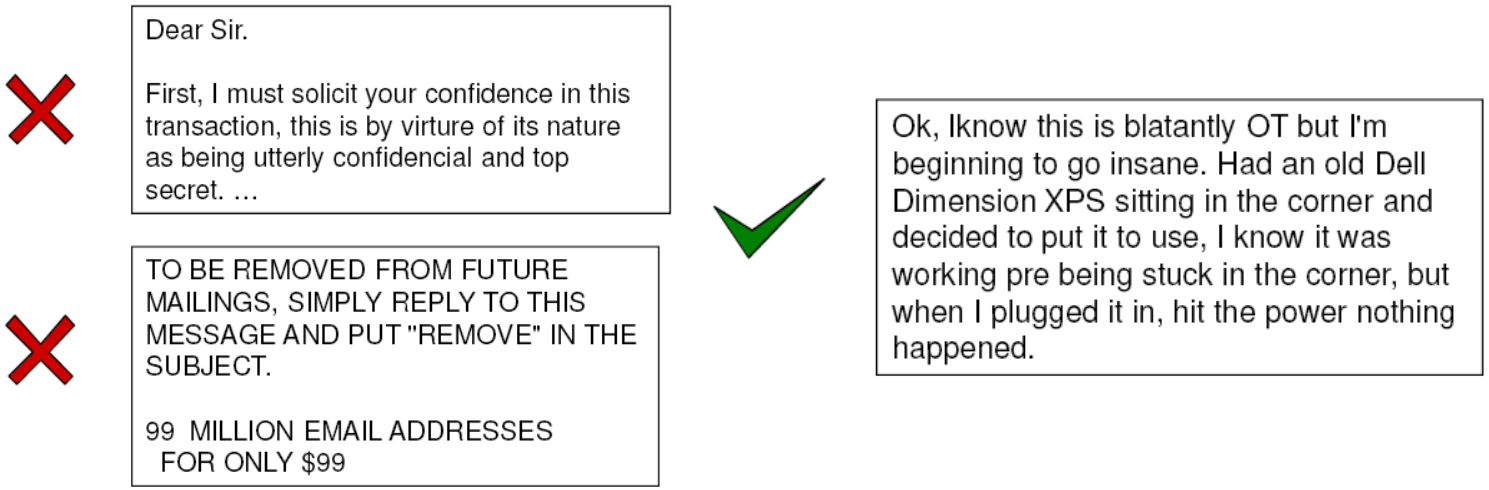
$$= \text{argmax} P(C = a) P(X_1 = x_1, \ldots, X_n = x_n | C = a)$$

$$\approx \text{argmax}\, P(C = a) P(x_1|a) P(x_2|a) \ldots P(x_n|a)$$

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)



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, …, 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 (w_1, …, 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 | 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/

Bayesian Inference and Bayesian Learning

- Bayes Rule
- Bayesian Inference
  - Misdiagnosis
  - The Bayesian “Decision”
  - The “Naïve Bayesian” Assumption
  - Bag of Words (BoW)
- Bayesian Learning
  - Maximum Likelihood estimation of parameters
  - Laplace Smoothing

Bayesian Learning

- 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

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

  - 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

Bayesian Learning

- 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, $X = [D_1, \ldots, D_m]$, each with its associated class label, $Y = [C_1, \ldots, C_m]$.
- 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:

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

Bayesian Learning

- 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, $D_i = [W_{1i}, \ldots, W_{ni}]$.
- 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:

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

Bayesian Learning

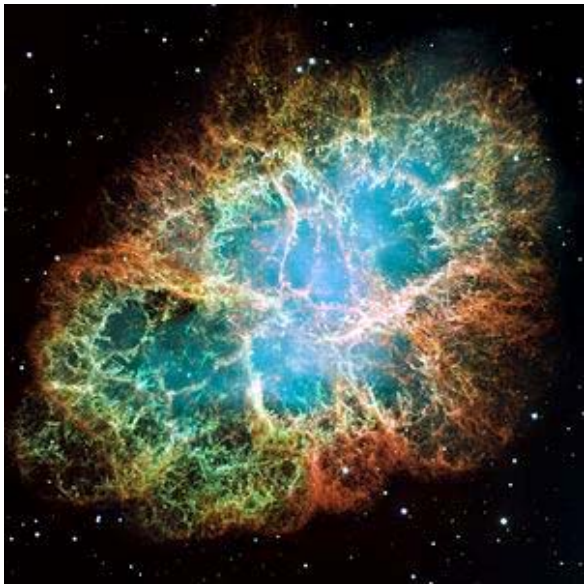
The data likelihood $P(X, Y)$ is maximized if we choose:

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

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

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 = $\frac{1}{1+100,000,001} = 0.000000009999998$

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

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

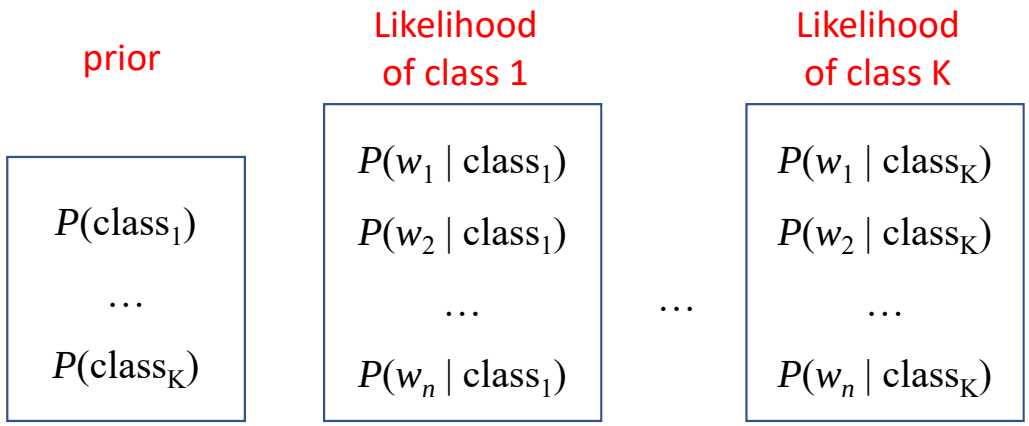
(V: total number of unique words)

Summary: Naïve Bayes for Document Classification

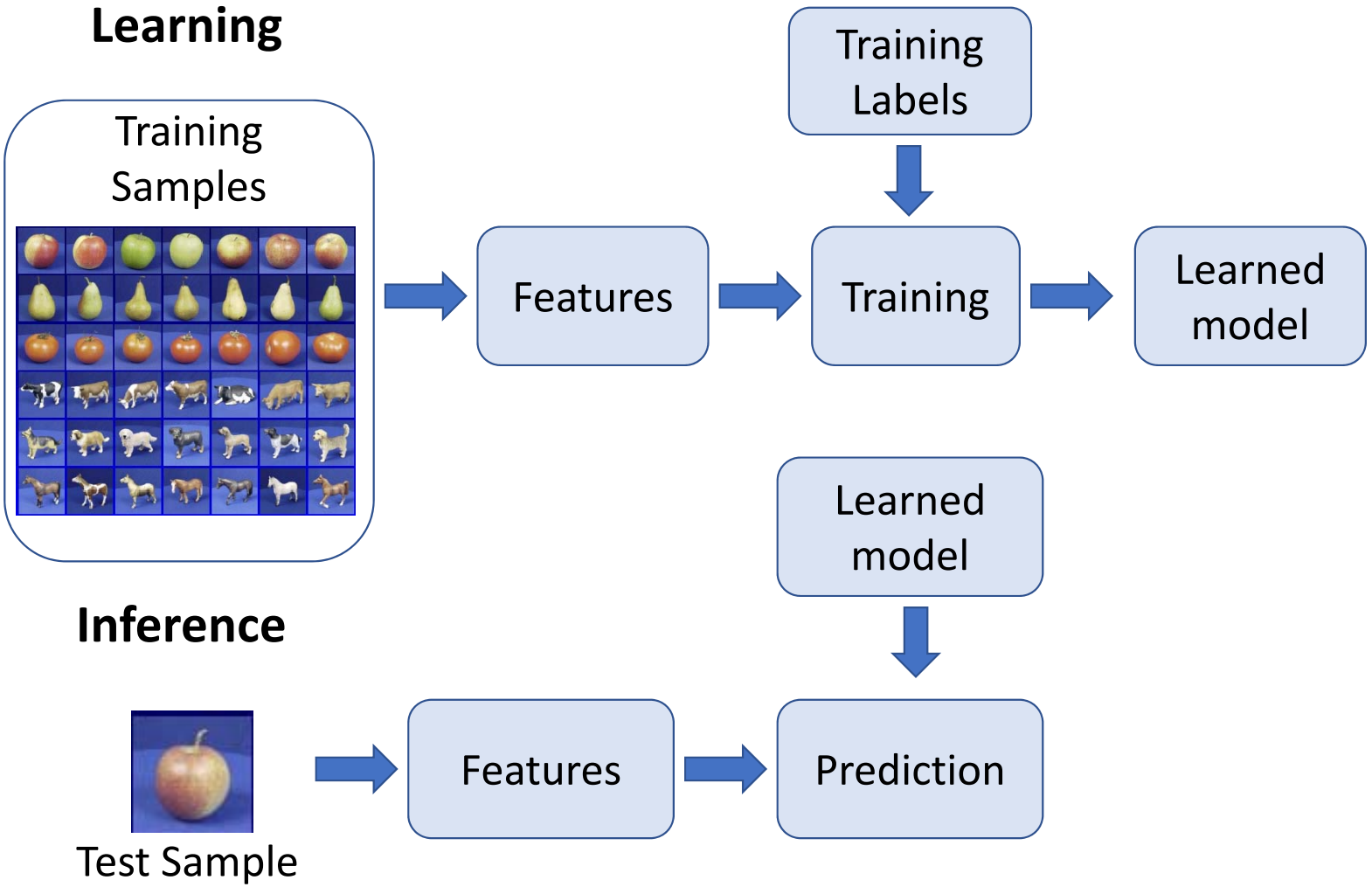
- Naïve Bayes model: assign the document to the class with the highest posterior

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

- Model parameters:



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* C based on the values of an observed *evidence variable* X
- **Inference problem:** given some observation $X = x$, what is $P(C \mid X=x)$?
- **Learning problem:** estimate the parameters of the probabilistic model $P(c \mid x)$ given a *training sample* $\{(x_1,c_1), \ldots, (x_n,c_n)\}$