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

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

Full joint distribution tables

x_1	x_2	x_3	…	y	p( )
0	0	0	…	0	*
0	0	0	…	1	*
1	0	0	…	0	*
1	0	0	…	1	*
0	1	0	…	0	*
0	1	0	…	1	*
			…		

Problem:

- all possible combination of features
- ~10,000 binary features
- Sample space size: 2^{10000}

Full joint distribution tables

x_1	x_2	x_3	…	y	p( )
0	0	0	…	0	*
0	0	0	…	1	*
1	0	0	…	0	*
1	0	0	…	1	*
0	1	0	…	0	*
0	1	0	…	1	*
			…		

- Storing a table of that size is impossible
- How are we supposed to learn/estimate each probability in the table?

Step 1: pick a model

$$p(features, label) = p(y)\prod_{j=1}^{m} p(x_i \mid y, x_1, \ldots, x_{i-1})$$

- So, far we have made NO assumptions about the data
- Model selection involves making assumptions about the data
- We did this before, e.g. assume the data is linearly separable
- The assumptions simplify the modeling task

e to the rescue

- ariables are independent if knowing the value of one
ng about the value of the other
- endent variables, knowing the value of one does not
bability distribution of the other variable (or the
any individual event)
- f the toss of a coin is independent of a roll of a die
- tea in England is independent of the whether or not
.

nt or dependent?

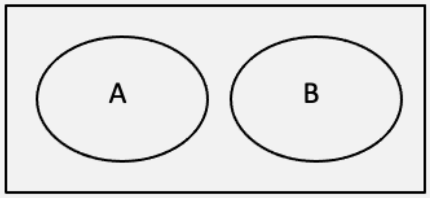
old and being allergic to cats

lon and driving habits

uccess as a mathematician

uccess as a basketball player

Independent random variables



If A and B are independent (written $A \perp\!\!\!\perp B$)

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

How does independence help us?

How does independence help us?

If A and B are independent

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

Independence

- Reduces the space needed to store distributions
- Reduces the complexity of the distribution
- Reduces the number of probabilities we need to estimate

ndependence

nts can become independent given certain other

weight

weight given genetics

onally independent given C

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

$= P(A|C)$

$= P(B|C)$

$) \neq P(A)P(B)$

Conditional Independence

- *X* is conditionally independent of Y given *Z* if the probability distribution governing *X* is independent of the value *of Y* given the value of *Z*:
- $P(X \mid Y, Z) = P(X \mid Z)$ that is

$$(\forall x_i, y_j, z_k) P(X = x_i \mid Y = y_j, Z = z_k) = P(X = x_i \mid Z = z_k)$$

Quick proof that independence is symmetric

- Assume: $P(X|Y, Z) = P(X|Y)$
- X and Z are independent given Y

$$P(Z \mid X,Y) = \frac{P(X,Y \mid Z)P(Z)}{P(X,Y)} \quad \text{(Bayes's Rule)}$$

$$= \frac{P(Y \mid Z)P(X \mid Y,Z)P(Z)}{P(X \mid Y)P(Y)} \quad \text{(Chain Rule)}$$

$$= \frac{P(Y \mid Z)P(X \mid Y)P(Z)}{P(X \mid Y)P(Y)} \quad \text{(By Assumption)}$$

$$= \frac{P(Y \mid Z)P(Z)}{P(Y)} = P(Z \mid Y) \quad \text{(Bayes's Rule)}$$

Naïve Bayes assumption

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

$$p(x_i \mid y, x_1, x_2, \ldots, x_{i-1}) = p(x_i \mid y)$$

What does this mean?

Naïve Bayes assumption

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

$$p(x_i \mid y, x_1, x_2, \ldots, x_{i-1}) = p(x_i \mid y)$$

Assumes ith feature is independent of the the other features *given the label*

Naïve Bayes assumption

$$p(x_i \mid y, x_1, x_2, \ldots, x_{i-1}) = p(x_i \mid y)$$

- Assumes ith feature is independent of the the other features *given the label*
- Assumes the probability of a word occurring in a review is independent of the other words *given the label*
- For example, the probability of the word "automatic" occurring is independent of whether or not "petrol" occurs given that the review is about "car"

Is this assumption true?

Naïve Bayes assumption

$$p(x_i \mid y, x_1, x_2, \ldots, x_{i-1}) = p(x_i \mid y)$$

- For most applications, this is not true!
- For example, the fact that “San” occurs will probably make it *more likely* that “Franciso” occurs (or take a compound phrase like “Penn State”)
- However, this is often a reasonable approximation:

$$p(x_i \mid y, x_1, x_2, \ldots, x_{i-1}) \approx p(x_i \mid y)$$

Naïve Bayes model

$$p(x_i \mid y, x_1, x_2, \ldots, x_{i-1}) = p(x_i \mid y)$$

$$= p(y) \prod_{i=1}^{m} p(x_i|y)$$

naïve bayes assumption

$p(x_i|y)$ is the probability of a particular feature value given the label

How do we model this?

- for binary features
- for count features
- for real valued features

Naïve Bayes Classifier

- How to learn $P(X|\omega_i)$?
- Naïve Bayes solution: Assume that the random variables in X are conditionally independent given the class.
- Result: Naïve Bayes classifier which performs optimally under certain assumptions
- A simple, practical learning algorithm grounded in Probability Theory

When to use

- Attributes that describe instances are likely to be conditionally independent given classification
- The data is insufficient to estimate all the probabilities reliably if we do not assume independence

Implications of Independence

- Suppose we have 5 Binary attributes and a binary class label
- Without independence, in order to specify the joint distribution, we need to specify a probability for each possible assignment of values to each variable resulting in a table of size $2^6=64$
- Suppose the features are independent given the class label – we only need $5(2\times2)=20$ entries
- The reduction in the number of probabilities to be estimated is even more striking when N, the number of attributes is large – from $O(2^N)$ to $O(N)$

Naive Bayes Classifier

Consider a discrete valued target function $f : \chi \rightarrow \Omega$
where an instance $X = (X_1, X_2 ... X_n) \in \chi$ is described
in termsof attribute values $X_1 = x_1,\ X_2 = x_2,\ ... X_n = x_n$
where $x_i \in Domain(X_i)$

$$\omega_{MAP} = \arg\max_{\omega_j \in \Omega} P(\omega_j \mid X_1 = x_1, X_2 = x_2 ... X_n = x_n)$$

$$= \arg\max_{\omega_j \in \Omega} \frac{P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid \omega_j) P(\omega_j)}{P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)}$$

$$= \arg\max_{\omega_j \in \Omega} P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid \omega_j) P(\omega_j)$$

ω_{MAP} is called the *maximum a posteriori* classification

Naive Bayes Classifier

$$\omega_{MAP} = \arg\max_{\omega_j \in \Omega} P(\omega_j \mid X_1 = x_1, X_2 = x_2 \ldots X_n = x_n)$$

$$= \arg\max_{\omega_j \in \Omega} P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \mid \omega_j) P(\omega_j)$$

If the attributes are *independent* given the class, we have

$$\omega_{MAP} = \arg\max_{\omega_j \in \Omega} \prod_{i=1}^{n} P\left(X_i = x_i \mid \omega_j\right) P(\omega_j)$$

$$= \omega_{NB}$$

$$= \arg\max_{\omega_j \in \Omega} P(\omega_j) \prod_{i=1}^{n} P\left(X_i = x_i \mid \omega_j\right)$$

Naive Bayes Learner

For each possible value ω_j of Ω,

$$\hat{P}(\Omega = \omega_j) \leftarrow Estimate(P(\Omega = \omega_j), D)$$

For each possible value a_{i_k} of X_i

$$\hat{P}(X_i = a_{i_k} \mid \omega_j) \leftarrow Estimate\left(P(X_i = a_{i_k} \mid \Omega = \omega_j), D\right)$$

Classify a new instance $X = (x_1, x_2, \ldots x_N)$

$$c(X) = \arg\max_{\omega_j \in \Omega} P(\omega_j) \prod_{i=1}^{n} P(X_i = x_i \mid \omega_j)$$

Estimate is a procedure for estimating the relevant probabilities from set of training examples

Learning Dating Preferences

Data samples – ordered 3-tuples of attribute values corresponding to

Height (tall, short)

Hair (dark, blonde, red)

Eye (blue, brown)

Classes: +, –

Training Data

Instance		Class label
I_1	(t, d, l)	+
I_2	(s, d, l)	+
I_3	(t, b, l)	–
I_4	(t, r, l)	–
I_5	(s, b, l)	–
I_6	(t, b, w)	+
I_7	(t, d, w)	+
I_8	(s, b, w)	+

Probabilities to estimate

$P(+) = 5/8$

$P(-) = 3/8$

$P(Height \mid c)$	t	s
$+$	3/5	2/5
$-$	2/3	1/3

$P(Hair \mid c)$	d	b	r
$+$	3/5	2/5	0
$-$	0	2/3	1/3

$P(Eye \mid c)$	l	w
$+$	2/5	3/5
$-$	1	0

Classify (*Height=t, Hair=b, eye=l*)

$$P(X \mid +) = (3/5)(2/5)(2/5) = (12/125)$$
$$P(X \mid -) = (2/3)(2/3)(1) = (4/9)$$
$$P(+|X) \propto P(+)P(X|+) = (5/8)(12/125) = 0.06$$
$$P(-|X) \propto P(-)P(X|-) = (3/8)(4/9) = 0.1667$$

Classify (*Height=t, Hair=r, eye=w*)

- Note the problem when a probability is 0
- Solution – Use Dirichlet Priors

$$P(x_i = a_{i_k}) = \frac{\alpha_{ik} + N_{ik}}{\sum_k \alpha_{ik} + \sum_k N_{ik}}$$

Modeling $p(x|y)$

Binary features:

$$p(x_i \mid y) = \begin{cases} \theta_i & \text{if } x_i = 1 \\ 1-\theta_i & \text{otherwise} \end{cases}$$

biased coin toss!

Other features:

Model using an appropriate distribution:

- Continuous values - Gaussian (i.e. normal) distribution
- Counts - Multinomial distribution (more on this later)
- …

- for discrete, we could simply do a much larger table, but often that doesn’t capture everything we want

Basic steps for probabilistic modeling

Probabilistic models

Step 1: pick a model

Which model do we use, i.e. how do we calculate p(*feature, label*)?

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

How do train the model, i.e. how to we we estimate the probabilities for the model?

Step 3 (optional): deal with overfitting

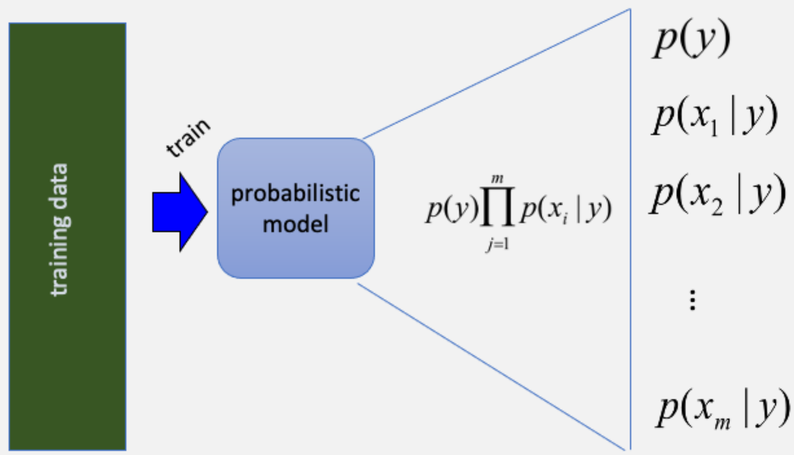
How do we deal with overfitting?

Obtaining probabilities

We've talked a lot about probabilities, but not where they come from

- How do we estimate $p(x_i|y)$ from training data?
- What is the probability of surviving the titanic?
- What is the probability that a review is about Toyota or BMW?

Obtaining probabilities



Estimating probabilities

What is the probability of a Macbook Air review?

We don't know!

We can ***estimate*** it based on data, though:

$$\frac{\text{number of reviews labeled Macbook Air}}{\text{total number of reviews}}$$

This is called the maximum likelihood estimation. Why?

Maximum Likelihood Estimation (MLE)

- Maximum likelihood estimation picks the values for the model parameters that *maximize the likelihood* of the training data
- You flip a coin 100 times. 60 times you get heads and 40 times you get tails.
- What is the MLE estimate for heads?

p(head) = 0.60 why?

Likelihood

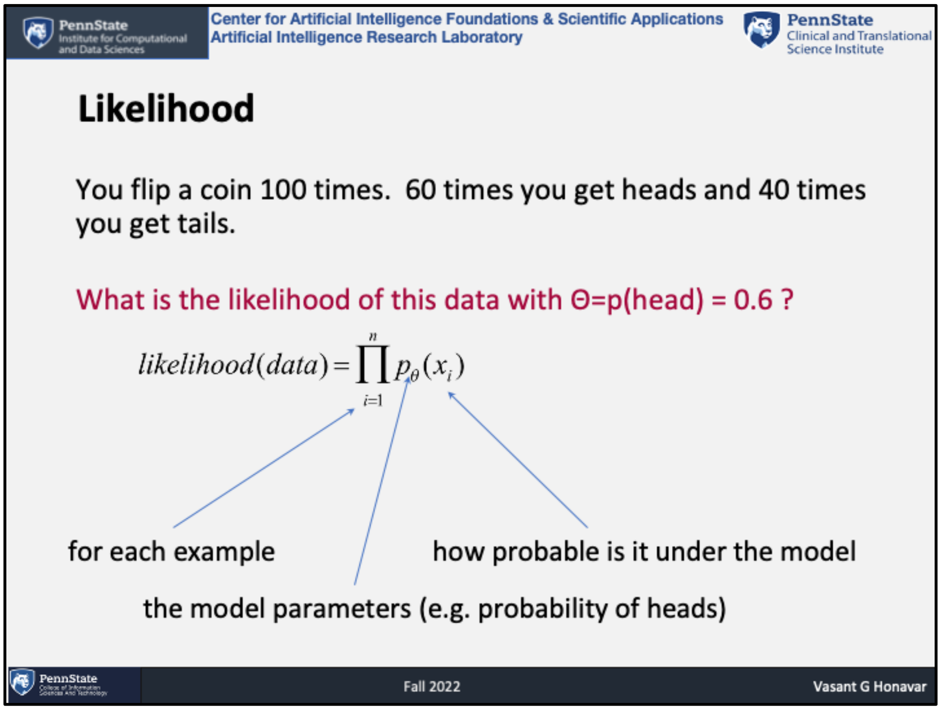
The likelihood of a data set is the probability that a particular model (i.e. one characterized by parameters, say θ) assigns to the data

$$likelihood(data) = \prod_{i=1}^{n} p_{\theta}(x_i)$$

for each example

how probable is it under the model

the model parameters (e.g. probability of heads)


Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory
PennState Institute for Computational and Data Sciences
PennState Clinical and Translational Science Institute
Likelihood
You flip a coin 100 times. 60 times you get heads and 40 times you get tails.
What is the likelihood of this data with Θ=p(head) = 0.6 ?
$likelihood(data) = \prod_{i=1}^{n} p_\theta(x_i)$
for each example
how probable is it under the model
the model parameters (e.g. probability of heads)
Fall 2022
Vasant G Honavar

Likelihood

You flip a coin 100 times. 60 times you get heads and 40 times you get tails.

$$0.60^{60} \times 0.40^{40} = 5.908465121038621e^{-30}$$

What is the likelihood of this data with Θ=p(head) = 0.6 ?

$$likelihood(data) = \prod_{i=1}^{n} p_{\theta}(x_i)$$

$$0.60^{60} \times 0.40^{40} = 5.908465121038621e^{-30}$$

60 heads with p(head) = 0.6

40 tails with p(tail) = 0.4

MLE example

Can we do any better?

$$likelihood(data) = \prod_i p(x_i)$$

$$0.60^{60} \times 0.40^{40} = 5.908465121038621e^{-30}$$

60 heads with $p(head)$ = 0.6   40 tails with $p(tail)$ = 0.4

What about $p(head)$ = 0.5?

MLE example

Can we do any better?

$$likelihood(data) = \prod_i p(x_i)$$

$$0.60^{60} \times 0.40^{40} = 5.908465121038621e^{-30}$$

60 heads with p(head) = 0.6

40 tails with p(tail) = 0.4

$$0.50^{60} \times 0.50^{40} = 7.888609052210118e^{-31}$$

60 heads with p(head) = 0.5

40 tails with p(tail) = 0.5

MLE example

Can we do any better?

$$likelihood(data) = \prod_i p(x_i)$$

$$0.60^{60} \times 0.40^{40} = 5.908465121038621e^{-30}$$

60 heads with p(head) = 0.6

40 tails with p(tail) = 0.4

What about p(head) = 0.7?

MLE example

Can we do any better?

$$likelihood(data) = \prod_i p(x_i)$$

$$0.60^{60} \times 0.40^{40} = 5.908465121038621e^{-30}$$

60 heads with p(head) = 0.6

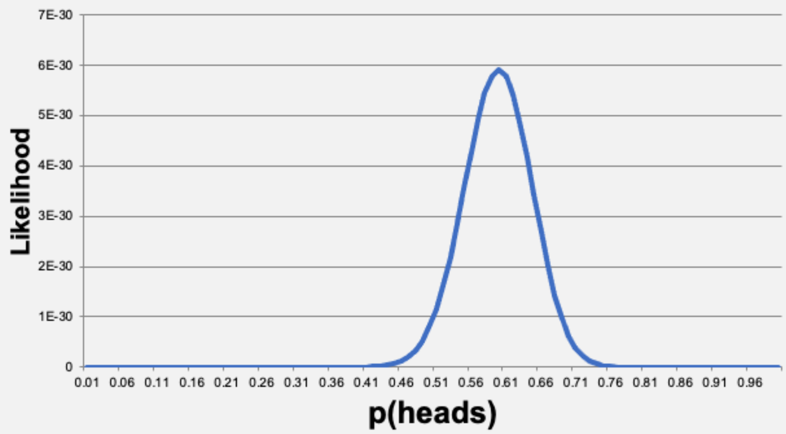
40 tails with p(tail) = 0.4

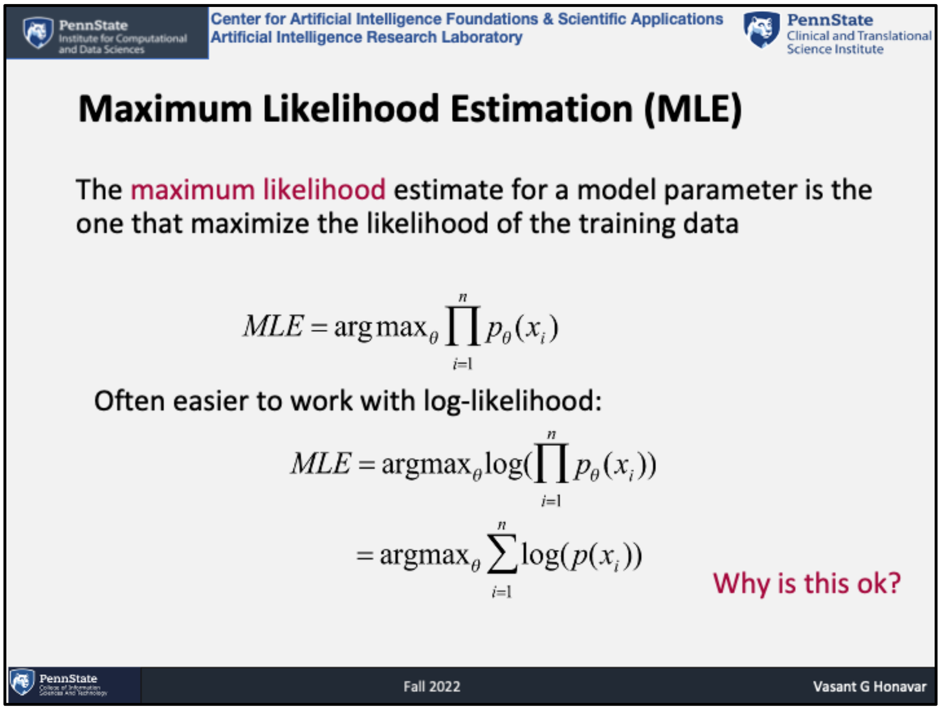
$$0.70^{60} \times 0.30^{40} = 6.176359828759916e^{-31}$$

60 heads with p(head) = 0.7

40 tails with p(tail) = 0.3

MLE Example


7E-30
6E-30
5E-30
4E-30
3E-30
2E-30
1E-30
0
Likelihood
0.01 0.06 0.11 0.16 0.21 0.26 0.31 0.36 0.41 0.46 0.51 0.56 0.61 0.66 0.71 0.76 0.81 0.86 0.91 0.96
p(heads)



- log is a strictly increasing function
- it just squishes values but does not change their order, so the max of likelihood is still the max of log-likelihood

Calculating MLE

The maximum likelihood estimate for a model parameter is the one that maximize the likelihood of the training data

$$MLE = \operatorname{argmax}_\theta \sum_{i=1}^{n} \log(p(x_i))$$

Given some training data, how do we calculate the MLE?

Training data: You flip a coin 100 times. 60 times you get heads and 40 times you get tails.

Calculating MLE

You flip a coin 100 times.

60 times you get heads and 40 times you get tails.

$$\log\text{-}likelihood = \sum_{i=1}^{n} \log(p(x_i))$$

$$= 60\log(p(heads)) + 40\log(p(tails))$$

$$= 60\log(\theta) + 40\log(1-\theta)$$

$$MLE = \arg\max_{\theta} 60\log(\theta) + 40\log(1-\theta)$$

How do we find the max?

Calculating MLE

- How do you find the maximum of a function of θ?

$$\frac{d}{d\theta} 60\log(\theta) + 40\log(1-\theta) = 0$$

$$\frac{60}{\theta} - \frac{40}{1-\theta} = 0$$

$$\frac{40}{1-\theta} = \frac{60}{\theta}$$

$$40\theta = 60 - 60\theta$$

$$100\theta = 60$$

$$\theta = \frac{60}{100}$$

Calculating MLE

You flip a coin n times.

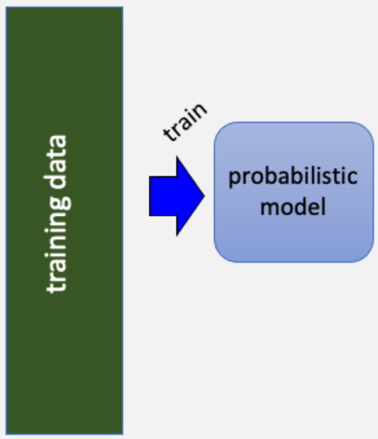
a times you get heads and *b* times you get tails.

$$\frac{d}{d\theta} a\log(\theta) + b\log(1-\theta) = 0$$

...

$$\theta = \frac{a}{a+b}$$

MLE estimation for Naïve Bayes



$$p(y) \prod_i^n p(x_i|y)$$

$$p(y) \qquad\qquad p(x_i \mid y)$$

What are the MLE estimates for these?

Center for Artificial Intelligence Foundations & Scientific Applications
Artificial Intelligence Research Laboratory

PennState Institute for Computational and Data Sciences

PennState Clinical and Translational Science Institute

Maximum likelihood estimates

$$p(y) = \frac{count(y)}{n}$$

$$\frac{\text{number of examples with label}}{\text{total number of examples}}$$

$$p(x_i \mid y) = \frac{count(x_i, y)}{count(y)}$$

$$\frac{\text{number of examples with the label with feature}}{\text{number of examples with label}}$$

What does training a NB model then involve?
How difficult is this to calculate?

PennState College of Information Sciences and Technology | Fall 2022 | Vasant G Honavar

- just involves iterating over the data and aggregating these counts!

Naïve Bayes classification

yellow, curved, no leaf, 6oz, banana → NB Model p(*features*, *label*) → 0.004

$$p(y)\prod_i^n p(x_i|\mathrm{y})$$

Given an unlabeled example:

predict the label

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

Probabilistic models

yellow, curved, no leaf, 6oz, banana

yellow, curved, no leaf, 6oz, apple

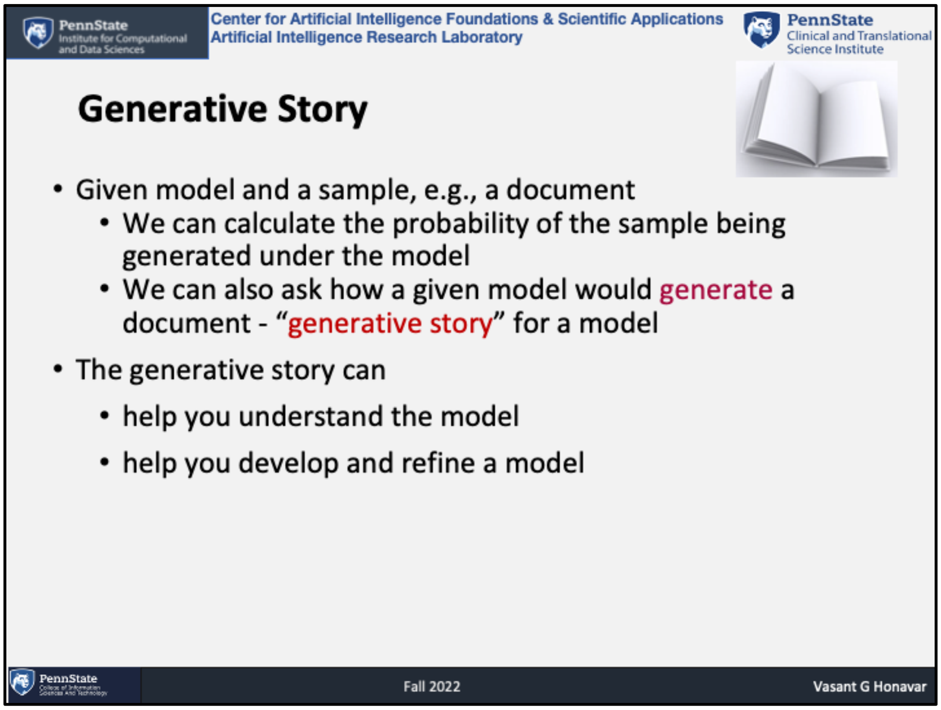
probabilistic model:

p(*features*, *label*)

pick largest

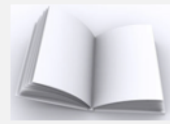
$$p(y)\prod_i^n p(x_i|\mathrm{y})$$

label = $\arg\max p(yj)\prod_i^n p(x_i|\mathrm{y_j})$



- although we don't generally "generate" a document from a model, it's often useful to look at the generative story of a model (i.e. how the model says a document was generate) to help us understand why the model assigns certain probabilities

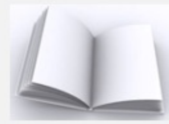
Naïve Bayes generative story



$$p(\omega\ )\prod_{i}^{n} P(x_i|\omega\ )$$

What is the generative story for the NB model?

NB generative story



$$p(\omega\ \ )\prod_{i}^{n} P(x_i|\omega\ \ )$$

1. Pick a label according to p(ω)
   - Roll a biased, m sided die
   - Set the class label that shows up
2. Given the class, for each of the n features feature:
   - Roll a biased die with as many sides as possible values of the feature.
   - Set the feature to the value that shows up