Slides adapted from Prof Carpuat and Duraiswami

CMSC 422 Introduction to Machine Learning

Lecture 16 A Probabilistic View of Machine Learning I & II

Furong Huang / furongh@umd.edu

UNIVERSITY OF MARYLAND

Today's topics

- Bayes rule review
- A probabilistic view of machine learning
  - Joint Distributions
  - Bayes optimal classifier
- Statistical Estimation
  - Maximum likelihood estimates
  - Derive relative frequency as the solution to a constrained optimization problem

Bayes Rule

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

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**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning*…

Exercise: Applying Bayes Rule

Consider the 2 random variables

- A = You have the flu
- B = You just coughed

Assume

- P(A) = 0.05
- P(B|A) = 0.8
- P(B|not A) = 0.2

What is P(A|B)?

Answer

- Via Logic
  - ✓ Assume 100 students – 5 have the flu. 80% (4) of the students who have the flu cough; 20% (19) of the students who don't have the flu cough; So the chance that you have the flu is 4/23
- Via Bayes Rule
  - ✓ $P(A|B)P(B)=P(B|A)P(A)$.
  - ✓ $P(B)=0.8*0.05+0.2*(1-0.05)=0.04+0.19=0.23$
  - ✓ $P(A|B)=0.8*0.05/0.23 =0.04/0.23=4/23$

Using a Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Using a Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Given the joint distribution, we can find the probability of any logical expression E involving these variables

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using a Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Given the joint distribution, we can make inferences

- E.g., P(Male | Poor)?
- Or P(Wealth | Gender, Hours)?

Recall: Machine Learning as Function Approximation

Problem setting

- Set of possible instances X
- Unknown target function $f: X \rightarrow Y$
- Set of function hypotheses $H = \{h \mid h: X \rightarrow Y\}$

Input

- Training examples $\{(x^{(1)}, y^{(1)}), \ldots (x^{(N)}, y^{(N)})\}$ of unknown target function f

Output

- Hypothesis $h \in H$ that best approximates target function f

Recall: Formal Definition of Binary Classification (from CIML)

TASK: BINARY CLASSIFICATION

Given:

1. An input space $\mathcal{X}$
2. An unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, +1\}$

Compute: A function f minimizing: $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[f(\boldsymbol{x}) \neq y\right]$

The Bayes Optimal Classifier

Assume we know the data generating distribution $\mathcal{D}$

We define the **Bayes Optimal classifier** as

$$f^{(\mathrm{BO})}(\hat{x}) = \arg\max_{\hat{y} \in \mathcal{Y}} \mathcal{D}(\hat{x}, \hat{y})$$

Theorem: Of all possible classifiers, the Bayes Optimal classifier achieves the smallest zero/one loss

Bayes error rate

- Defined as the error rate of the Bayes optimal classifier
- Best error rate we can ever hope to achieve under zero/one loss

The Bayes Optimal Classifier

Assume we know the data generating distribution $\mathcal{D}$

We define the **Bayes Optimal classifier** as

$$f^{(\mathrm{BO})}(\hat{x}) = \arg\max_{\hat{y}} \mathcal{D}(\hat{x}, \hat{y})$$

Theo fier
achi

Baye
D
Bo

If we had access to $\mathcal{D}$,
Finding an optimal classifier would be trivial!

we don't have access to $\mathcal{D}$
So let's try to estimate it instead!

What does "training" mean in probabilistic settings?

- Training = estimating $\mathcal{D}$ from a finite training set
  - We typically assume that $\mathcal{D}$ comes from a specific family of probability distributions
  - e.g., Bernouilli, Gaussian, etc
  - Learning means inferring parameters of that distributions
  - e.g., mean and covariance of the Gaussian

Training assumption: training examples are iid

- **Independently and Identically distributed**
  - i.e. as we draw a sequence of examples from $\mathcal{D}$, the n-th draw is independent from the previous n-1 sample

- This assumption is usually false!
  - But sufficiently close to true to be useful

How can we estimate the joint probability distribution from data?

What are the challenges?

Maximum Likelihood Estimation

- Find the parameters that maximize the probability of the data

- Example: how to model a biased coin?

  (on board)

Maximum Likelihood Estimates

X=1 X=0

P(X=1) = θ

P(X=0) = 1-θ

(Bernoulli)

Each coin flip yields a Boolean value for X

X ~ Bernouilli: $P(X) = \theta^X(1-\theta)^{1-X}$

Given a data set D of iid flips, which contains α_1 ones and α_0 zeros

$$P_\theta(D) = \theta^{\alpha_1}(1-\theta)^{\alpha_0}$$

$$\hat{\theta}_{MLE} = argmax_\theta \; P_\theta(D) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

Maximum Likelihood Estimation

- Example: how to model a k-sided die? (on board)

What we know so far…

- Bayes rule
- A probabilistic view of machine learning
  - If we know the data generating distribution, we can define the Bayes optimal classifier
  - Under iid assumption
- How to estimate a probability distribution from data?
  - Maximum likelihood estimation

Let's learn a classifier by learning P(Y|X)

Goal: learn a classifier P(Y|X)

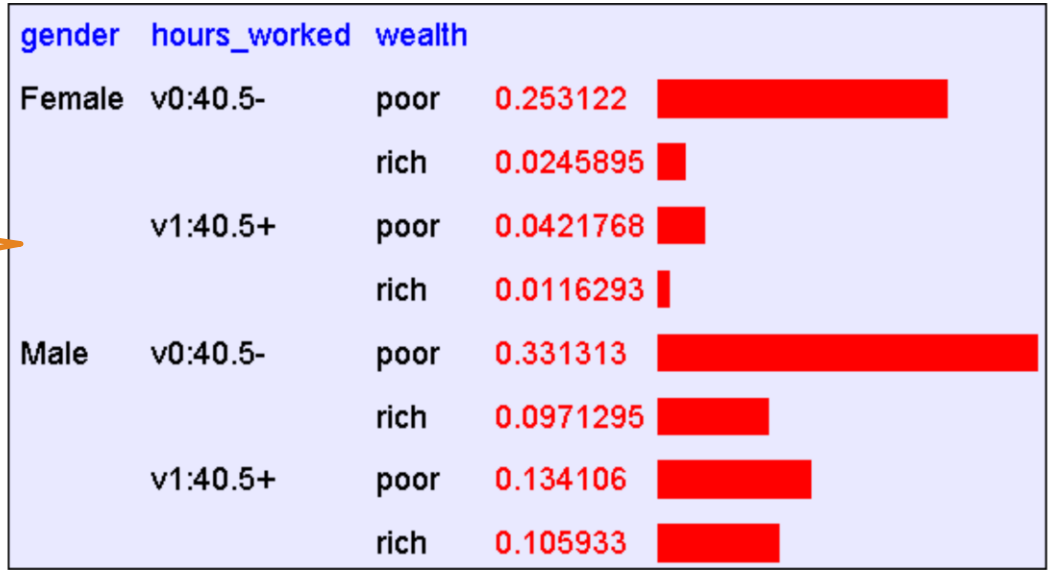
Prediction:

- Given an example x
- Predict $\hat{y} = argmax_y \, P(Y = y \,|X = x)$

Parameters for P(X,Y) vs. P(Y|X)

Y = Wealth
X = <Gender, Hours_worked>

Joint probability distribution P(X,Y)



gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Conditional probability distribution P(Y|X)

Gender	HrsWorked	P(rich | G,HW)	P(poor | G,HW)
F	<40.5	.09	.91
F	>40.5	.21	.79
M	<40.5	.23	.77
M	>40.5	.38	.62

How many parameters do we need to learn?

Suppose $X = < X_1, X_2, \ldots X_d >$

where X_i and Y are Boolean random variables

Q: How many parameters do we need to estimate $P(Y|X_1, X_2, \ldots X_d)$?

A: Too many to estimate P(Y|X) directly from data!

Naïve Bayes Assumption

Naïve Bayes assumes

$$P(X_1, X_2, \ldots X_d|Y) = \prod_{i=1}^{d} P(X_i|Y)$$

i.e., that X_i and X_j are **conditionally independent** given Y, for all $i \neq j$

Conditional Independence

Definition:

X is conditionally independent of Y given Z

if **P(X|Y,Z) = P(X|Z)**

Recall that X is independent of Y if P(X|Y)=P(X)

Naïve Bayes classifier

$$\hat{y} = argmax_y\, P(Y = y \,|X = x)$$
$$= argmax_y P(Y = y)P(X = x \,|Y = y)$$
$$= argmax_y P(Y = y) \prod_{i=1}^{d} P(X_i = x_i \,|Y = y)$$

Bayes rule

\+ Conditional independence assumption

How many parameters do we need to learn?

To describe P(Y)?

To describe $P(X = < X_1, X_2, \ldots X_d > | Y)$

Without conditional independence assumption?

With conditional independence assumption?

(Suppose all random variables are Boolean)

Training a Naïve Bayes classifier

Let's assume discrete Xi and Y 

TrainNaïveBayes (Data)

for each value y_k of Y

estimate $\pi_k = P(Y = y_k)$

$$\frac{\#\ examples\ for\ which\ Y = y_k}{\#\ examples}$$

for each value x_{ij} of X_i

estimate $\theta_{ijk} = P(X_i = x_{ij} \mid Y = y_k)$

$$\frac{\#\ examples\ for\ which\ X_i = x_{ij}\ and\ Y = y_k}{\#\ examples\ for\ which\ Y = y_k}$$

Naïve Bayes Wrap-up

An easy to implement classifier, that performs well in practice

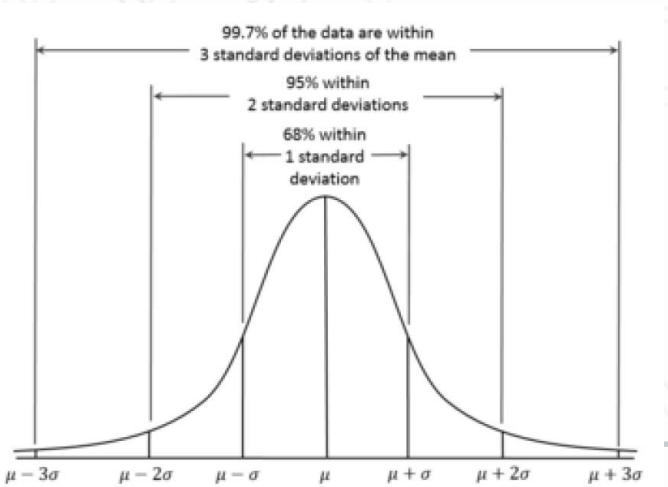
What is the decision boundary of a Naïve Bayes classifier?

Naïve Bayes Properties

Naïve Bayes is a linear classifier

- See CIML for example of computation of Log Likelihood Ratio

Choice of probability distribution is a form of inductive bias



Generative Stories

Probabilistic models tell a fictional story explaining how our training data was created

Example of a generative story for a multiclass classification task with continuous features

For each example $n = 1 \dots N$:

(a) Choose a label $y_n \sim \mathcal{D}isc(\boldsymbol{\theta})$

(b) For each feature $d = 1 \dots D$:

i. Choose feature value $x_{n,d} \sim \mathcal{N}or(\mu_{y_n,d}, \sigma^2_{y_n,d})$

From the Generative Story to the Likelihood Function

For each example $n = 1 \dots N$:

(a) Choose a label $y_n \sim \mathcal{D}isc(\boldsymbol{\theta})$

(b) For each feature $d = 1 \dots D$:

i. Choose feature value $x_{n,d} \sim \mathcal{N}or(\mu_{y_n,d}, \sigma^2_{y_n,d})$

$$p(D) = \overbrace{\prod_n \underbrace{\theta_{y_n}}_{\text{choose label}} \underbrace{\prod_d \frac{1}{\sqrt{2\pi\sigma^2_{y_n,d}}} \exp\left[-\frac{1}{2\sigma^2_{y_n,d}}(x_{n,d} - \mu_{y_n,d})^2\right]}_{\text{choose feature value}}}^{\text{for each example}}$$

for each feature

What you should know

The Naïve Bayes classifier

- Conditional independence assumption
- How to train it?
- How to make predictions?
- How does it relate to other classifiers we know?

Fundamental Machine Learning concepts

- iid assumption
- Bayes optimal classifier
- Maximum Likelihood estimation
- Generative story

UNIVERSITY OF
MARYLAND

Furong Huang
4124 IRB, College Park, MD 20740
301.405.8010 / furongh@umd.edu