A Probabilistic View of Machine Learning (2/2)

CMSC 422

MARINE CARPUAT
marine@cs.umd.edu

Some slides based on material by Tom Mitchell

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

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

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 Estimates

K sided die

$\forall\, k, P(X = k) = \theta_k$

A constrained optimization problem

$$\hat{\theta}_{MLE} \quad = \quad argmax_{\theta} \sum_{k=1}^{K} x_k \log(\theta_k)$$

$$with \qquad \sum_{k=1}^{K} \theta_k = 1$$

Can be solved using e.g., Lagrange Multipliers (on board)

$$\hat{\theta}_k = \frac{x_k}{\sum_{i=1}^{K} x_i}$$

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 estimate?

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

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

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 $\mathbf{P(X|Y,Z) = P(X|Z)}$

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

Naïve Bayes classifier

$$\hat{y} = argmax_y\, P(Y = y \mid X = x)$$

$$= argmax_y P(Y = y) P(X = x \mid Y = y)$$

$$= argmax_y P(Y = y) \prod_{i=1}^{d} P(X_i = x_i \mid Y = y)$$

Bayes rule

+ Conditional independence assumption

How many parameters do we need to estimate?

- To describe P(Y)?

- To describe $P(X = < X_1, X_2, \dots X_d > | Y)$
  - 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 $\theta_{ijk} = P\left(X_i = x_{ij} \,\middle|\, Y = y_k\right)$

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

Naïve Bayes Properties

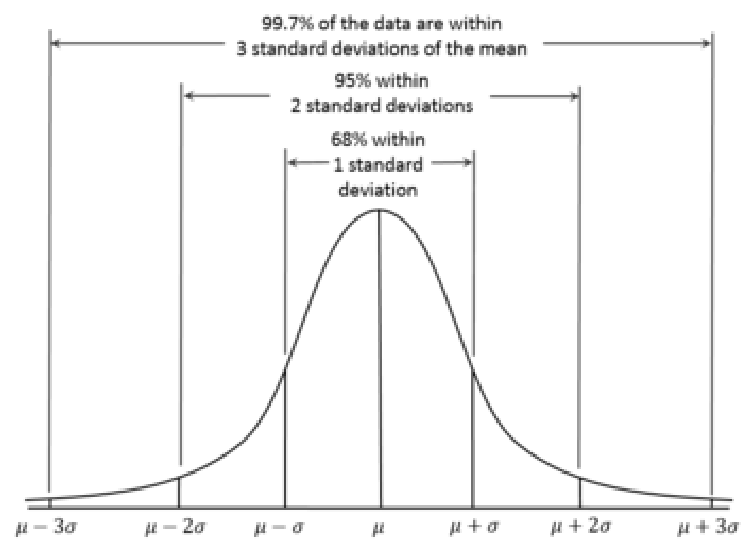
- A simple, easy to implement classifier, that performs well in practice
- Subtleties
  - Often the Xi are not really conditionally independent
  - What if the Maximum Likelihood estimate for P(Xi|Y) is zero?

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

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

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 & tools
  - iid assumption
  - Bayes optimal classifier
  - Maximum likelihood estimation
  - Lagrange multipliers