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

Each coin flip yields a Boolean value for $X$

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

Given a data set $D$ of iid flips, which contains $\alpha_1$ ones and $\alpha_0$ zeros

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

$$\hat{\theta}_{MLE} = \arg\max_\theta P_\theta(D) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$
Maximum Likelihood Estimates

K-sided die: each die roll yields an integer between 1 and K

Given a data set D of iid rolls, where side $i$ is observed $x_i$ times:

$$P_\theta(D) = \prod_{i=1}^{K} \theta_i^{x_i}$$

$$\hat{\theta}_{i,MLE} = \frac{x_i}{\sum_{i=1}^{K} x_i}$$
Let’s learn a classifier by learning $P(Y|X)$

- **Goal:** learn a classifier $P(Y|X)$

- **Prediction:**
  - Given an example $x$
  - Predict $\hat{y} = \arg\max_y P(Y = y \mid X = x)$
Parameters for $P(X,Y)$ vs. $P(Y|X)$

$Y = \text{Wealth}$

$X = \langle \text{Gender, Hours\_worked} \rangle$

Joint probability distribution $P(X,Y)$

| gender | hours\_worked | wealth | $P(Y|X)$ |
|---------|----------------|--------|----------|
| 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 | $P(\text{rich} | G,HW)$ | $P(\text{poor} | G,HW)$ |
|--------|-----------|-----------|-----------|
| 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 $X = < X_1, X_2, ... X_d >$

where $X_i$ and $Y$ are Boolean random variables

Q: How many parameters do we need to estimate $P(Y|X_1, X_2, ... X_d)$?
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 \textbf{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 $P(X|Y) = P(Y)$
Naïve Bayes classifier

\[
\hat{y} = \arg\max_y P(Y = y \mid X = x) \\
= \arg\max_y P(Y = y) P(X = x \mid Y = y) \\
= \arg\max_y P(Y = y) \prod_{i=1}^{d} P(X_i = x_i \mid Y = y)
\]

Bayes rule

+ Conditional independence assumption
How many parameters do we need to estimate?

• To describe $P(Y)$?
• To describe $P(X = < X_1, X_2, ... X_d > | Y)$?
  – Without conditional independence assumption?
  – With conditional independence assumption?

(Suppose all random variables are Boolean)
Training a Naïve Bayes classifier

Let’s assume discrete $X_i$ and $Y$

**TrainNaïveBayes (Data)**

1. For each value $y_k$ of $Y$:
   - Estimate $\pi_k = P(Y = y_k)$
2. For each value $x_{ij}$ of $X_i$:
   - Estimate $\theta_{ijk} = P(X_i = x_{ij} \mid Y = y_k)$

$\theta_{ijk} = \frac{\text{# examples for which } X_i = x_{ij} \text{ and } Y = y_k}{\text{# examples for which } Y = y_k}$
Naïve Bayes Wrap-up

• An easy to implement classifier, that performs well in practice

• Subtleties
  – Often the $X_i$ are not really conditionally independent
  – What if the Maximum Likelihood estimate for $P(X_i = x_i \mid Y = y)$ 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 \( n = 1 \ldots N \):

  (a) Choose a label \( y_n \sim Disc(\theta) \)

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

    i. Choose feature value \( x_{n,d} \sim Nor(\mu_{y_n,d}, \sigma_{y_n,d}^2) \)
From the Generative Story to the Likelihood Function

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

(a) Choose a label $y_n \sim Disc(\theta)$

(b) For each feature $d = 1 \ldots D$:
   i. Choose feature value $x_{n,d} \sim Nor(\mu_{y_{n,d}}, \sigma^2_{y_{n,d}})$

\[
p(D) = \prod_n \theta_{y_n} \prod_d \frac{1}{\sqrt{2\pi\sigma^2_{y_{n,d}}}} \exp \left[ -\frac{1}{2\sigma^2_{y_{n,d}}} (x_{n,d} - \mu_{y_{n,d}})^2 \right]
\]

for each example

choose label

choose feature value

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