A Probabilistic View of Machine Learning, Naïve Bayes

CMSC 422
SOHEIL FEIZI
sfeizi@cs.umd.edu

Slides adapted from MARINE CARPUAT
Logistics

• Midterm next lecture!

• Programming assignment will be available on Thursday
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

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$  Bayes’ rule

we call $P(A)$ the “prior”

and $P(A|B)$ the “posterior”


...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
  
  \( A = \text{You have the flu} \)
  
  \( B = \text{You just coughed} \)

- Assume
  
  \( P(A) = 0.05 \)
  
  \( P(B|A) = 0.8 \)
  
  \( P(B|\text{not } A) = 0.2 \)

- What is \( P(A|B) \)?
Using a Joint Distribution

<table>
<thead>
<tr>
<th>gender</th>
<th>hours_worked</th>
<th>wealth</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.253122</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0245895</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
<td>0.0421768</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0116293</td>
</tr>
<tr>
<td>Male</td>
<td>v0:40.5-</td>
<td>poor</td>
<td>0.331313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0971295</td>
</tr>
<tr>
<td></td>
<td>v1:40.5+</td>
<td>poor</td>
<td>0.134106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.105933</td>
</tr>
</tbody>
</table>
Using a Joint Distribution

- 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

Given the joint distribution, we can make inferences
- E.g., $P(\text{Male} | \text{Poor})$?
- Or $P(\text{Wealth} | \text{Gender}, \text{Hours})$?
Recall: Machine Learning as Function Approximation

Problem setting
• Set of possible instances $X$
• Unknown target function $f: X \rightarrow Y$
• Set of function hypotheses $H = \{h | h: X \rightarrow Y\}$

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: $\mathbb{E}_{(x,y) \sim \mathcal{D}}[f(x) \neq y]$
The Bayes Optimal Classifier

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

• We define the **Bayes Optimal classifier** as

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

• The 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

• 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?
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

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

$$\hat{\theta}_{MLE} = \arg \max_{\theta} P_{\theta}(D) = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$
Let’s learn a classifier by learning $P(Y|X)$

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

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

$Y = \text{Wealth}$

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

**Joint probability distribution $P(X,Y)$**

<table>
<thead>
<tr>
<th>Gender</th>
<th>HrsWorked</th>
<th>Wealth</th>
<th>$P(X,Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>$&lt;40.5$</td>
<td>poor</td>
<td>0.253122</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0245895</td>
</tr>
<tr>
<td></td>
<td>$&gt;40.5$</td>
<td>poor</td>
<td>0.0421768</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0116293</td>
</tr>
<tr>
<td>Male</td>
<td>$&lt;40.5$</td>
<td>poor</td>
<td>0.331313</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.0971295</td>
</tr>
<tr>
<td></td>
<td>$&gt;40.5$</td>
<td>poor</td>
<td>0.134106</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rich</td>
<td>0.105933</td>
</tr>
</tbody>
</table>

**Conditional probability distribution $P(Y|X)$**

| Gender | HrsWorked | $P(Y|X)$ |
|--------|-----------|---------|
| F      | $<40.5$   | .09     |
|        | $>40.5$   | .21     |
| M      | $<40.5$   | .23     |
|        | $>40.5$   | .38     |
How many parameters do we need to learn?

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 \( P(Y|X_1, X_2, ... X_d) \)?

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

Naïve Bayes assumes

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

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 $P(X|Y)=P(Y)$
Naïve Bayes classifier

\[ \hat{y} = \arg \max_y P(Y = y | X = x) \]

\[ = \arg \max_y P(Y = y)P(X = x | Y = y) \]

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

Bayes rule

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

• To describe $P(Y)$?

• To describe $P(X = < X_1, X_2, ... X_d > | Y)$
  – 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)

for each value $y_k$ of $Y$

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

for each value $x_{ij}$ of $X_i$

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

\[
\frac{\text{# examples for which } Y = y_k}{\text{# examples}}
\]

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

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

• Subtleties
  – Often the Xi are not really conditionally independent
  – What if the Maximum Likelihood estimate for \( P(X_i | Y) \) is zero?