EECS 391: Introduction to AI

Soumya Ray
Website: http://engr.case.edu/ray_soumya/eecs391_sp16/
Email: sray@case.edu
Office: Olin 516
Office hours: M 12:30-2pm (W 12:30-2 this week) or by appointment
Announcements

• HW4, PA3 due today
• Quiz Tuesday
  – Bring abaci/log tables/slide rules/calculators
Today

• Machine Learning, Probabilistic Classification
  – Ch 18.1-2, 20.1, 20.2.[1,2,4]
### Example

<table>
<thead>
<tr>
<th></th>
<th>Has-fur?</th>
<th>Long-Teeth?</th>
<th>Scary?</th>
<th>Lion?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Animal\textsubscript{1}</strong></td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td><strong>Animal\textsubscript{2}</strong></td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Animal\textsubscript{3}</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Attribute-value representation**

\(X\)

**Binary Class Label (\(y\))**

(assigned by target concept/teacher)
Probabilistic Classification

- Given examples as before, one way to solve the learning problem is:
  - Treat the features \(X\) and the class label \(Y\) as random variables
  - Construct a probabilistic model of \(p(X=x, Y=y)\)
  - Given a new example, for which \(X\) is known and has value \(x\), calculate \(p(Y=y|X=x)\)
    - Return the \(y\) with highest probability
The Naïve Bayes Classifier

\[ p(Y = y) \]

\[ p(X_i = x_i \mid Y = y) \]
The Naïve Bayes Classifier

• Simple probabilistic classifier for discrete data

\[ p_{X,Y}(x, y) = \prod_{i} p(X_i = x_i | Y = y) p(Y = y) \]

Naïve Bayes assumption: Attributes are conditionally independent given the class.

Naïve Bayes parameters: Instead of storing probabilities for each atomic event, we will only store these conditional probabilities and use this formula to calculate the probability for an atomic event (example).
Classification with naïve Bayes

<table>
<thead>
<tr>
<th>Has-fur?</th>
<th>Long-Teeth?</th>
<th>Scary?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal₁</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

\[
p(\text{Has-fur}=\text{Yes} | \text{Lion})=0.5, \quad p(\text{Has-fur}=\text{Yes} | \text{Not-Lion})=0.1
\]
\[
p(\text{Long-Teeth}=\text{Yes} | \text{Lion})=0.9, \quad p(\text{Long-Teeth}=\text{Yes} | \text{Not-Lion})=0.5
\]
\[
p(\text{Scary}=\text{Yes} | \text{Lion})=0.8, \quad p(\text{Scary}=\text{Yes} | \text{Not-Lion})=0.5
\]
\[
p(\text{Lion})=0.1
\]

\[
P(\text{Animal1 AND Lion})=0.1 \times 0.5 \times 0.1 \times 0.2
\]
\[
P(\text{Animal1 AND Not-Lion})=0.9 \times 0.1 \times 0.5 \times 0.5
\]
Learning a Naïve Bayes classifier

• Given a set of observations:

<table>
<thead>
<tr>
<th>Has-fur?</th>
<th>Long-Teeth?</th>
<th>Scary?</th>
<th>Lion?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal₁</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Animal₂</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Animal₃</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

• Learn parameters $p(X_i=x_i|Y=y)$ and $p(Y=y)$
We will use a method called “Maximum Likelihood Estimation”
Bayes Rule for Concept Learning

• Suppose we are given a set of examples $D$ and we are considering a hypothesis space $H$ (e.g. the space all naive Bayes classifiers) to try to find the target concept.

• The posterior probability of any hypothesis $h$ in $H$ is given by Bayes Rule:

$$\Pr(h \mid D) = \frac{\Pr(D \mid h) \Pr(h)}{\Pr(D)}$$
MAP Hypothesis

• Given: training sample $D$ and hypothesis class $H$

• Do: Return the most probable hypothesis given the data---the maximum a posteriori (MAP) hypothesis

$$h_{MAP} = \arg \max_{h \in H} \Pr(h \mid D)$$

$$= \arg \max_{h \in H} \frac{\Pr(D \mid h) \Pr(h)}{\Pr(D)}$$

$$= \arg \max_{h \in H} \Pr(D \mid h) \Pr(h)$$
ML Hypothesis

• If every hypothesis in $H$ has equal prior probability, only the first term matters.
• This gives the maximum likelihood (ML) hypothesis:

$$h_{ML} = \arg \max_{h \in H} \Pr(D \mid h)$$

• Let us use this to find the maximum likelihood naïve Bayes classifier for a given training set.
Maximum Likelihood Estimation

- For naïve Bayes, a hypothesis is the vector of parameters, one for each of $p(X_i=x_i|Y=y)$ and $P(Y=y)$

- Assume $X_i$ is 0/1 and $Y$ is 0/1
  - Then $p(X_i=1|Y=1)$ is a parameter, call it $\theta_{i1}$
  - There’s another parameter for $p(X_i=1|Y=0)$, $\theta_{i0}$
  - Finally there are two parameters for $p(Y=y)$, $\theta_y$ ($\theta_0$ and $\theta_1$—these sum to 1)
Maximum Likelihood Estimation

$$h_{ML} = \arg \max_{h \in H} p(D \mid h)$$

$$p(D \mid h) = p(\{x_d, y_d\}_{d=1}^{m} \mid \{\theta_{i0}, \theta_{i1}\}_{i=1}^{n}, \theta_y)$$

$$= \prod_{d=1}^{m} p(x_d, y_d \mid \{\theta_{i0}, \theta_{i1}\}_{i=1}^{n}, \theta_y)$$

$$= \prod_{d=1}^{m} \prod_{i=1}^{n} p(X_i = x_i \mid Y = y_d; \{\theta_{i0}, \theta_{i1}\}, \theta_y) p(Y = y_d)$$

$$= \prod_{d=1}^{m} \prod_{i=1}^{n} p(X_i = x_i \mid Y = y_d; \{\theta_{i0}, \theta_{i1}\}, \theta_y) \theta_{y_d}$$
<table>
<thead>
<tr>
<th>Animal</th>
<th>Has-fur? (f1)</th>
<th>Long-Teeth? (f2)</th>
<th>Scary? (f3)</th>
<th>Lion? (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal₁</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Animal₂</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Animal₃</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

\[
p(D \mid h) = \prod_{d=1}^{m} \prod_{i=1}^{n} p(X_i = x_i \mid Y = y_d; \{\theta_{i0}, \theta_{i1}\}) \theta_{y_d}
\]

\[
= \left[ \theta_{10} (1 - \theta_{20})(1 - \theta_{30}) \theta_{0} \right] \times \\
\left[ (1 - \theta_{10}) \theta_{20} \theta_{30} \theta_{0} \right] \times \\
\left[ \theta_{11} \theta_{21} \theta_{31} \theta_{1} \right]
\]

Let \( N_0 \) be the number of examples with \( Y=0 \) and suppose \( d_i \) of those have \( f_i=Yes \).
\[ p(D | h) = \prod_{d=1}^{m} \prod_{i=1}^{n} p(X_i = x_i | Y = y_d ; \theta_{i0}, \theta_{i1}) \theta_{yd} \]

For \( Y = 0 \), \( p(D | h) = \prod_{i=1}^{n} \theta_{i0}^{d_i} (1 - \theta_{i0})^{N_0 - d_i} \theta_{0}^{N_0} \)

\[ \hat{\theta}_{k0} = \arg \max_{\theta_{k0}} \theta_{k0}^{d_k} (1 - \theta_{k0})^{N_0 - d_k} = L(\theta_{k0}) \]

\[ LL(\theta_{k0}) = d_k \log \theta_{k0} + (N_0 - d_k) \log(1 - \theta_{k0}) \]

\[ \frac{\partial LL}{\partial \theta_{k0}} = \frac{d_k}{\theta_{k0}} - \frac{(N_0 - d_k)}{(1 - \theta_{k0})} = 0 \]

or \( d_k - d_k \theta_{k0} = N_0 \cdot \theta_{k0} - d_k \theta_{k0} \)

or \( d_k = N_0 \cdot \theta_{k0} \)

or \( \hat{\theta}_{k0} = \frac{d_k}{N_0} \)

Number of examples with \( Y=0 \)

Likelihood function/Log likelihood function

Fraction of observed \( Y=0 \) examples where \( X_k=1 \)!
Naïve Bayes Parameter MLEs

\[
\hat{p}(X_i = 1 | Y = 1) = \frac{\text{# observed examples with } X_i = 1 \text{ and } Y = 1}{\text{# observed examples with } Y = 1}
\]

\[
p(X_i = 1 | Y = 1) = \frac{p(X_i = 1, Y = 1)}{p(Y = 1)}
\]

\[
\hat{p}(Y = 1) = \frac{\text{# observed examples with } Y = 1}{\text{# observed examples}}
\]
## Example

<table>
<thead>
<tr>
<th>Has-fur?</th>
<th>Long-Teeth?</th>
<th>Scary?</th>
<th>Lion?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal₁</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Animal₂</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Animal₃</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

\[
p(\text{Has-fur}=\text{Yes} \mid \text{Lion})=?, \quad p(\text{Has-fur}=\text{Yes} \mid \text{Not-Lion})=? \\
p(\text{Long-Teeth}=\text{Yes} \mid \text{Lion})=?, \quad p(\text{Long-Teeth}=\text{Yes} \mid \text{Not-Lion})=? \\
p(\text{Scary}=\text{Yes} \mid \text{Lion})=?, \quad p(\text{Scary}=\text{Yes} \mid \text{Not-Lion})=? \\
p(\text{Lion})=?
\]
Smoothing probability estimates

• What happens if a certain value for a variable is not in our set of examples, for a certain class?
  – Suppose we’re trying to classify lions and we’ve never seen a lion cub, so $p_{MLE}(Scary = false | Lion) = 0$
  – When we see a cub, its probability of being a lion will be zero by our Naïve Bayes formula, even if it has long teeth and fur
  – It’s a good idea to “smooth” our probability estimates to avoid this
**m-Estimates**

\[
p(X_i = x_i \mid Y = y) = \frac{(\text{#examples with } X_i = x_i \text{ and } Y = y) + mp}{(\text{#examples with } Y = y) + m}
\]

- \(p\) is our prior estimate of the probability
- \(m\) is called “Equivalent Sample Size” which determines the importance of \(p\) relative to the observations
- If variable has \(v\) values, the specific case of \(m=v\), \(p=1/v\) is called Laplace smoothing
Applications

• Machine learning methods are widely used
  – Most things you do online use some sort of ML in the back end for something

• Perhaps more interestingly, naïve Bayes is probably running on your computer at the moment
Application: Email Spam Filtering

• Naïve Bayes has been used very successfully to categorize documents
  – Is this document about “sports” or “finance”?  
  – Is this email “spam” or “ham”? 

• Given a vocabulary, each attribute $X_i$ is the presence/absence of word $i$ in the document
  – Ignores word order  
  – “Bag-of-words” approach
Email Spam Filtering with Naïve Bayes

- Smoothed parameter estimates
  \[ p(\text{word}_k \text{ present} | Y = \text{spam}) = \]
  \[ \frac{(#\text{emails with word}_k \text{ present and } Y = \text{spam}) + 1}{(#\text{emails with } Y = \text{spam}) + 2} \]

- Variations used by most major email clients/commercial spam filters e.g. DSPAM, SpamAssassin, SpamBayes, Bogofilter, ASSP
  - Read article on “Bayesian spam filtering” in wikipedia
Summary

• We learned about:
  – Machine Learning
  – Probabilistic Classification
  – Naïve Bayes
  – Maximum likelihood estimation
  – Smoothing parameters

• Next: Sequential Decision Making (part 4)
  – Ch 17, 21