EECS 391: Introduction to AI

Soumya Ray

Website: http://vorlon.case.edu/~sray/eecs391_sp12/index.html

Email: sray@case.edu

Office: Olin 516

Office hours: Tue 2:30-4:00 or by appointment

Binary Classification

	Has-fur?	Long-Teeth?	Scary?	*Lion?*
Animal$_1$	Yes	No	No	No
Animal$_2$	No	Yes	Yes	No
Animal$_3$	Yes	Yes	Yes	Yes

Attribute-value representation ($\mathbf{X}$)

Binary Class Label (y) (assigned by target concept/ teacher)

Naïve Bayes

- Simple probabilistic classifier for discrete data

$$p_{\mathbf{X},Y}(\mathbf{x}, y) = p(\mathbf{X} = \mathbf{x} \,|\, Y = y)\, p(Y = y)$$
$$= p(x_1, \ldots, x_n \,|\, Y = y)\, p(Y = y)$$
$$= \prod_i p(X_i = x_i \,|\, Y = y)\, p(Y = y)$$

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).

Probability Estimation for Naïve Bayes

- Given a set of observations:

	Has-fur?	Long-Teeth?	Scary?	*Lion?*
Animal$_1$	Yes	No	No	No
Animal$_2$	No	Yes	Yes	No
Animal$_3$	Yes	Yes	Yes	Yes

- Estimate 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 hypotheses space H
- Each hypothesis could have "caused" the data we observe
- The posterior probability of any hypothesis h in H is given by Bayes Rule:

Likelihood Prior

Posterior $\Pr(h \mid D) = \dfrac{\Pr(D \mid h)\Pr(h)}{\Pr(D)}$

Evidence

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

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

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 binary (0/1)
  - Then $p(X_i=1|Y=1)$ is a parameter, call it θ_{i1}
  - There's another parameter for $p(X_i=1|Y=0)$, θ_{i0}
  - Finally there's a parameter for $p(Y=y)$, θ_y

Maximum Likelihood Estimation

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

$$p(D \mid h) = p(\{\mathbf{x}_d, y_d\}_{d=1\ldots m} \mid \{\theta_{i0}, \theta_{i1}\}_{i=1\ldots n}, \theta_y)$$

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

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

	Has-fur? (f1)	Long-Teeth? (f2)	Scary? (f3)	*Lion?* (Y)
Animal$_1$	Yes	No	No	No
Animal$_2$	No	Yes	Yes	No
Animal$_3$	Yes	Yes	Yes	Yes

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

$$= \left[\theta_{10}(1-\theta_{20})(1-\theta_{30})\theta_{y_0} \right] \times$$

$$\left[(1-\theta_{10})\theta_{20}\theta_{30}\theta_{y_0} \right] \times$$

$$\left[\theta_{11}\theta_{21}\theta_{31}\theta_{y_1} \right]$$

Let *pos* be the number of examples with $Y=1$ and d_i of those have $f_i=Yes$

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

Number of examples with $Y=1$

$$\text{For } Y = 1,\ p(D \mid h) = \prod_{i=1}^{n} \theta_{i1}^{d_i} (1-\theta_{i1})^{pos-d_i} \theta_{y_1}^{pos}$$

$$\hat{\theta}_{i1} = \arg\max_{\theta_{i1}} \prod_{i=1}^{n} \theta_{i1}^{d_i} (1-\theta_{i1})^{pos-d_i} \theta_{y_1}^{pos} = L(\theta_{i1})$$

$$LL(\theta_{i1}) = pos \log \theta_{y_1} + \sum_{i=1}^{n} d_i \log \theta_{i1} + (pos - d_i) \log(1-\theta_{i1})$$

$$\frac{\partial LL}{\partial \theta_{i1}} = \frac{d_i}{\theta_{i1}} - \frac{(pos - d_i)}{(1-\theta_{i1})} = 0$$

$$\text{or } d_i - d_i\theta_{i1} = pos \cdot \theta_{i1} - d_i\theta_{i1}$$

$$\text{or } d_i = pos \cdot \theta_{i1}$$

$$\text{or } \theta_{i1} = d_i / pos$$

Let *pos* be the number of examples with $Y=1$ and d_i of those have $f_i=Yes$

Fraction of observed $Y=1$ examples where $X_i=1$!

Naïve Bayes Parameter MLEs

$$\hat{p}(X_i = 1 \mid Y = 1) = \frac{\#\text{ observed examples with } X_i = 1 \text{ and } Y = 1}{\#\text{ observed examples with } Y = 1}$$

$$p(X_i = 1 \mid Y = 1) = \frac{p(X_i = 1, Y = 1)}{p(Y = 1)}$$

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

Example

	Has-fur?	Long-Teeth?	Scary?	*Lion?*
$Animal_1$	Yes	No	No	No
$Animal_2$	No	Yes	Yes	No
$Animal_3$	Yes	Yes	Yes	Yes

p(Has-fur=Yes|Lion)=?, p(Has-fur=Yes|Not-Lion)=?
p(Long-Teeth=Yes|Lion)=?, p(Long-Teeth=Yes|Not-Lion)=?
p(Scary=Yes|Lion)=?, p(Scary=Yes|Not-Lion)=?
p(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 $\hat{p}(Scary = false \mid Lion) = 0$
  - 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

Email Spam Filtering with Naïve Bayes

- 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

$$p(word_k \text{ present} \mid Y = spam) = \frac{(\#\text{emails with } word_k \text{ present and } Y = spam) + 1}{(\#\text{emails with } Y = spam) + 2}$$

Bernoulli distribution

- 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:
  - Maximum likelihood estimation
  - Smoothing parameters
- Next: Artificial Neural Networks