CSDS 440: Machine Learning

Soumya Ray (he/him, sray@case.edu)

Olin 516

Office hours T, Th 11:15-11:45 or by appointment

Announcements

- Last homework and programming assignments posted
- Projects posted---start now
- Quizzes: 2 of 4

Two Approaches to Probabilistic Classification

- Generative approaches model the joint distribution $p(\mathbf{x},y)$

- Discriminative approaches model the conditional distribution $p(y|\mathbf{x})$

Naïve Bayes

- Simplest generative classifier for discrete data

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

Naïve Bayes assumption: Attributes are conditionally independent given the class

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

Example

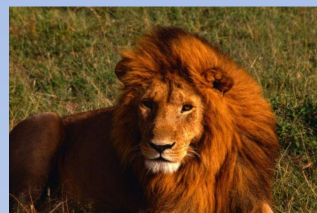
	Has-fur?	Long-Teeth?	Scary?	*Lion?*
Animal_1	Yes	No	No	No
Animal_2	No	Yes	Yes	No
Animal_3	Yes	Yes	Yes	Yes

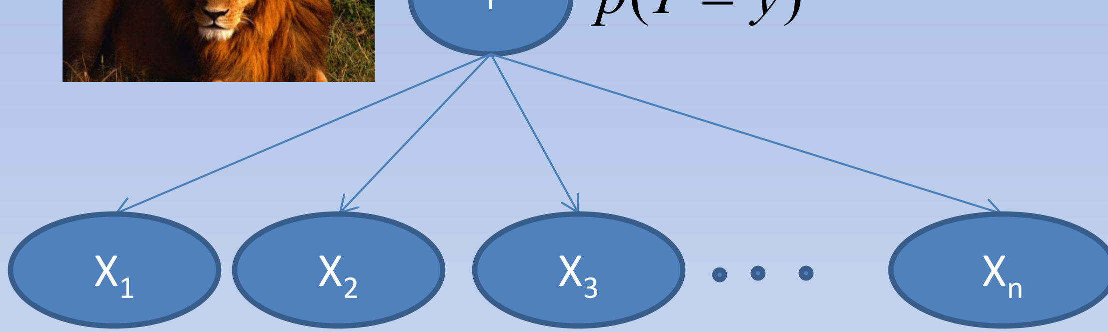
Naïve Bayes parameters:
p(Lion), p(Has-fur|Lion), p(Not-Has-fur|Lion), p(Long-Teeth|Lion), p(Not-Long-Teeth|Lion), p(Scary|Lion), p(Not-Scary|Lion)
p(Not-Lion), p(Has-fur|Not-Lion), p(Not-Has-fur|Not-Lion), p(Long-Teeth|Not-Lion), p(Not-Long-Teeth|Not-Lion), p(Scary|Not-Lion), p(Not-Scary|Not-Lion)

How many parameters?

- Two for $p(Y=y)$
- One each for $p(X_i=x_i|Y=y)$
  - Suppose X_i is Boolean
- $2(2n+1)$ total---much better than 2^{n+1}
  - Of these, need to estimate only $2n+1$

Aside: A Graphical View of Naïve Bayes





$$p(Y = y)$$

$$p(X_i = x_i \mid Y = y)$$

The class label Y "causes" each attribute X_i to have a certain value, independently of each other attribute.

Probabilistic Graphical Model (CSDS 491)
Bayesian Network (CSDS 391/491)

Classification with Naïve Bayes

- For a new example, calculate $p(\mathbf{X}=\mathbf{x}, Y=\text{“}positive\text{”})$ and $p(\mathbf{X}=\mathbf{x}, Y=\text{“}negative\text{”})$ and choose whichever is greater

$$p(\mathbf{X}=\mathbf{x}, Y=pos)=$$
$$\prod_i p(X_i = x_i \mid Y = pos)\, p(Y = pos)$$

Example

	Has-fur?	Long-Teeth?	Scary?
Animal_1	Yes	No	No

p(Has-fur=Yes|Lion)=0.5, p(Has-fur=Yes|Not-Lion)=0.1
p(Long-Teeth=Yes|Lion)=0.9, p(Long-Teeth=Yes|Not-Lion)=0.5
p(Scary=Yes|Lion)=0.8, p(Scary=Yes|Not-Lion)=0.5
p(Lion)=0.1

$p(\text{Animal}_1, \text{Lion})=0.1*0.2*0.1*0.5=0.001$
$p(\text{Animal}_1, \text{Not-Lion})=0.9*0.5*0.5*0.1=0.0225$
So Animal_1 is more likely to not be a lion.

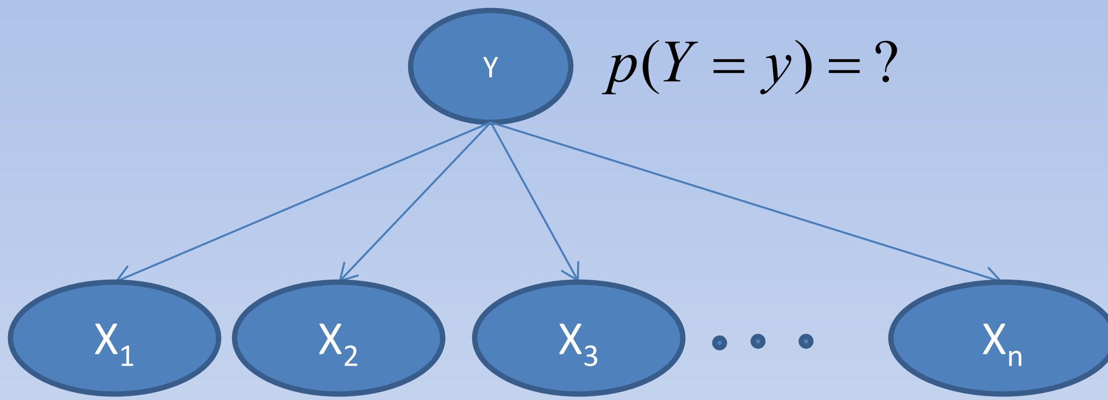
Learning a Naïve Bayes classifier

- 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 $p(Y=y)$

Estimating parameters



$$p(Y=y)=?$$

$$p(X_i=x_i \mid Y=y)=?$$

We will use Maximum Likelihood Estimation

Bayes Rule for Learning

- Suppose we are given a set of examples D and we are considering a set of candidate hypotheses H
- The posterior probability of any hypothesis h in H is given by Bayes Rule:

$$\boxed{\text{Posterior}} \quad \Pr(h \mid D) = \frac{\overbrace{\Pr(D \mid h)}^{\text{Likelihood}}\,\overbrace{\Pr(h)}^{\text{Prior}}}{\underbrace{\Pr(D)}_{\text{Evidence}}}$$

MAP Hypothesis

- Given: examples D and set of hypotheses H
- Do: Return the most probable hypothesis given the data---the maximum a posteriori (MAP) hypothesis

$$
\begin{aligned}
h_{MAP} &= \arg\max_{h \in H} \Pr(h \mid D) \\
&= \arg\max_{h \in H} \frac{\Pr(D \mid h)\Pr(h)}{\Pr(D)} \\
&= \arg\max_{h \in H} \Pr(D \mid h)\Pr(h)
\end{aligned}
$$

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 0/1 and Y is 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 are two parameters for $p(Y=y)$, θ_y (θ_0 and θ_1—these sum to 1)

Maximum Likelihood Estimation

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

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

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

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

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

	Has-fur? (f1)	Long-Teeth? (f2)	Scary? (f3)	*Lion?* (Y)
Animal$_1$	1	0	0	0
Animal$_2$	0	1	1	0
Animal$_3$	1	1	1	1

$$p(D \mid h) = \left[\theta_{10}(1-\theta_{20})(1-\theta_{30})\theta_0\right]\times$$

$$\left[(1-\theta_{10})\theta_{20}\theta_{30}\theta_0\right]\times\left[\theta_{11}\theta_{21}\theta_{31}\theta_1\right]$$

$$=\theta_{10}^1(1-\theta_{10})^1\theta_{20}^1(1-\theta_{20})^1\theta_{30}^1(1-\theta_{30})^1\theta_0^2\times$$

$$\theta_{11}^1(1-\theta_{11})^0\theta_{21}^1(1-\theta_{21})^0\theta_{31}^1(1-\theta_{31})^0\theta_1^1$$

Let N_1 be the number of examples with $Y=1$ and suppose p_i of those have $X_i=1$
Let N_0 be the number of examples with $Y=0$ and suppose d_i of those have $X_i=1$

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

$$=\prod_{i=1}^{n}\theta_{i1}^{p_i}(1-\theta_{i1})^{N_1-p_i}\,\theta_1^{N_1}\prod_{i=1}^{n}\theta_{i0}^{d_i}(1-\theta_{i0})^{N_0-d_i}\,\theta_0^{N_0}$$

Number of examples with $Y=0$

Number of $Y=0$ examples with $f_i=1$

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

Likelihood function

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

Log likelihood function

$$\frac{\partial LL}{\partial\theta_{k0}}=\frac{d_k}{\theta_{k0}}-\frac{(N_0-d_k)}{(1-\theta_{k0})}=0,\text{so }\frac{d_k}{\theta_{k0}}=\frac{(N_0-d_k)}{(1-\theta_{k0})}$$

$$\text{or }d_k-d_k\theta_{k0}=N_0\cdot\theta_{k0}-d_k\theta_{k0}$$

$$\text{or }d_k=N_0\cdot\theta_{k0}$$

$$\text{or }\hat{\theta}_{k0}={d_k}/{N_0}$$

Fraction of observed $Y=0$ examples where $X_k=1$!

Naïve Bayes Parameter MLEs

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

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

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

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

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

- 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

Nominal Attributes

- Need to estimate parameters $p(X_i{=}v_k|Y{=}y)$

- Can use maximum likelihood estimates:

$$p(X_i = v_k \mid Y = y) = \frac{p(X_i = v_k \wedge Y = y)}{p(Y = y)}$$

$$= \frac{\#\text{examples with } X_i = v_k \text{ and } Y = y}{\#\text{examples with } Y = y}$$

Continuous Attributes

- If X_i is a continuous attribute, can model $p(X_i|y)$ as a Gaussian distribution ("Gaussian naïve Bayes")

$$p(X_i \mid y) \sim N(\mu_{i|y}, \sigma_{i|y})$$

- MLEs

$$\hat{\mu}_i = \frac{\sum_{k \in examples} x_{ik} I(y_k = y)}{\sum_{k \in examples} I(y_k = y)}$$

$$\hat{\sigma}_i^2 = \frac{\sum_{k \in examples} (x_{ik} - \hat{\mu}_i)^2 I(y_k = y)}{\sum_{k \in examples} I(y_k = y)}$$

Naïve Bayes Geometry

- What does the decision surface of the naïve Bayes classifier look like?
- An example is classified positive iff

$$p(\mathbf{x},y=1) > p(\mathbf{x},y=0)$$

$$\frac{p(\mathbf{x}, y=1)}{p(\mathbf{x}, y=0)} > 1$$

$$\frac{\prod_i p(x_i \mid y=1)p(y=1)}{\prod_i p(x_i \mid y=0)p(y=0)} > 1$$

Naïve Bayes Geometry

- Classify an example as positive if

$$\frac{\prod_i p(x_i \mid y=1)p(y=1)}{\prod_i p(x_i \mid y=0)p(y=0)} > 1$$

$$\ln \frac{\prod_i p(x_i \mid y=1)p(y=1)}{\prod_i p(x_i \mid y=0)p(y=0)} > 0$$

$$\ln \frac{p(y=1)}{p(y=0)} + \sum_i \ln\left(\frac{p(x_i \mid y=1)}{p(x_i \mid y=0)}\right) > 0$$

Naïve Bayes Geometry

$$\ln \frac{p(y=1)}{p(y=0)} + \sum_i \ln\left(\frac{p(x_i \mid y=1)}{p(x_i \mid y=0)}\right) > 0$$

Indicator function

$$\ln \frac{p(y=1)}{p(y=0)} + \sum_i \sum_v \ln\left(\frac{p(X_i = v \mid y=1)}{p(X_i = v \mid y=0)}\right) I(X_i = v) > 0$$

$$(b_1 - b_0) + \sum_{i,v} (w_{iv1} - w_{iv0}) I(X_i = v) > 0,$$

$$b_1 = \ln p(y=1), w_{iv1} = \ln p(X_i = v \mid y=1)$$

$$b_0 = \ln p(y=0), w_{iv0} = \ln p(X_i = v \mid y=0)$$

So Naïve Bayes implements a **linear decision boundary, but with a logarithmic parameterization**