CSDS 440: Machine Learning

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

Olin 516

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

Zoom Link

Announcements

- Quiz 3 Thursday
  - Topics up to and including Optimization

Recap

- Naïve Bayes factorizes the j____ distribution as the product of _____. This assumes that _______.
- To infer the label of a new example, we ________.
- We estimate parameters for probabilistic models using M_____ L_____ E______.
- Bayes Rule for concept learning says that the p_____ is equal to the l_____ times the p____ divided by the e_____.
- Maximizing the LHS gives us the ____ hypothesis.
- If we assume that all hypotheses have e__ p___, we get the ____ hypothesis.
- To apply MLE, we first write down the L_____ f____. We then optimize it with respect to the m_____ p_____.

Today

- Probabilistic Machine Learning

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 $X_i=1$

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

Number of examples with $Y=0$

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

Number of $Y=0$ 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**

Why does Naïve Bayes work well?

- Very simplistic independence assumptions
  - Everyone knows that these assumptions are nearly always wrong
  - But, paradoxically, often works well in practice

- Why?
  - Works well for *classification*, but not so great at *density estimation*
  - Most probabilities end up near 0/1 (ask for paper)

Logistic Regression

- Simplest Discriminative model
- Models log odds as a linear function

$$\log \frac{p(Y=1\,|\,\mathbf{x})}{p(Y=-1\,|\,\mathbf{x})} = \mathbf{w}\bullet\mathbf{x} + b$$

$$p(Y=1\,|\,\mathbf{x}) = \left[1 - p(Y=1\,|\,\mathbf{x})\right] e^{(\mathbf{w}\bullet\mathbf{x}+b)}$$

$$p(Y=1\,|\,\mathbf{x})(1 + e^{(\mathbf{w}\bullet\mathbf{x}+b)}) = e^{(\mathbf{w}\bullet\mathbf{x}+b)}$$

$$p(Y=1\,|\,\mathbf{x}) = \frac{e^{(\mathbf{w}\bullet\mathbf{x}+b)}}{1 + e^{(\mathbf{w}\bullet\mathbf{x}+b)}} = \frac{1}{1 + e^{-(\mathbf{w}\bullet\mathbf{x}+b)}}$$

Classification with LR

- LR directly specifies $p(Y=1|\mathbf{x})$, compute and check if greater than 1/2

Estimating parameters

- Use MLE, optimize *conditional* log likelihood of the data

$$\mathbf{w}, b = \arg\max \prod_i p(Y_i = y_i \mid \mathbf{x}_i)$$

Conditional Likelihood

Conditional Log Likelihood

$$= \arg\max \sum_{i \in pos} \log p(Y_i = 1 \mid \mathbf{x}_i) + \sum_{i \in neg} \log p(Y_i = -1 \mid \mathbf{x}_i)$$

$$= \arg\max \sum_{i \in pos} \log\left(\frac{1}{1+e^{-\mathbf{w}\bullet\mathbf{x}+b}}\right) + \sum_{i \in neg} \log\left(1 - \frac{1}{1+e^{-\mathbf{w}\bullet\mathbf{x}+b}}\right)$$

Overfitting control

- Can include a term for overfitting control:

$$\mathbf{w}, b = \arg\max \sum_{i \in pos} \log\left(\frac{1}{1+e^{-\mathbf{w}\bullet\mathbf{x}+b}}\right) + \sum_{i \in neg} \log\left(1 - \frac{1}{1+e^{-\mathbf{w}\bullet\mathbf{x}+b}}\right)$$

$$\mathbf{w}, b = \arg\min \frac{1}{2}\left\|\mathbf{w}\right\|^2 + C\left[\begin{array}{l} \displaystyle\sum_{i \in pos} -\log\left(\frac{1}{1+e^{-\mathbf{w}\bullet\mathbf{x}+b}}\right) \\ \displaystyle +\sum_{i \in neg} -\log\left(1 - \frac{1}{1+e^{-\mathbf{w}\bullet\mathbf{x}+b}}\right) \end{array}\right]$$

Negative Conditional Log Likelihood

Estimating parameters

- Can use gradient descent, Newton methods etc

- Very robust method, works extremely well in many practical situations, very easy to code

Logistic Regression Geometry

- Classify as positive iff:

$$\frac{p(Y=1\,|\,\mathbf{x})}{p(Y=-1\,|\,\mathbf{x})} > 1$$

$$\text{or if } \log\frac{p(Y=1\,|\,\mathbf{x})}{p(Y=-1\,|\,\mathbf{x})} > 0$$

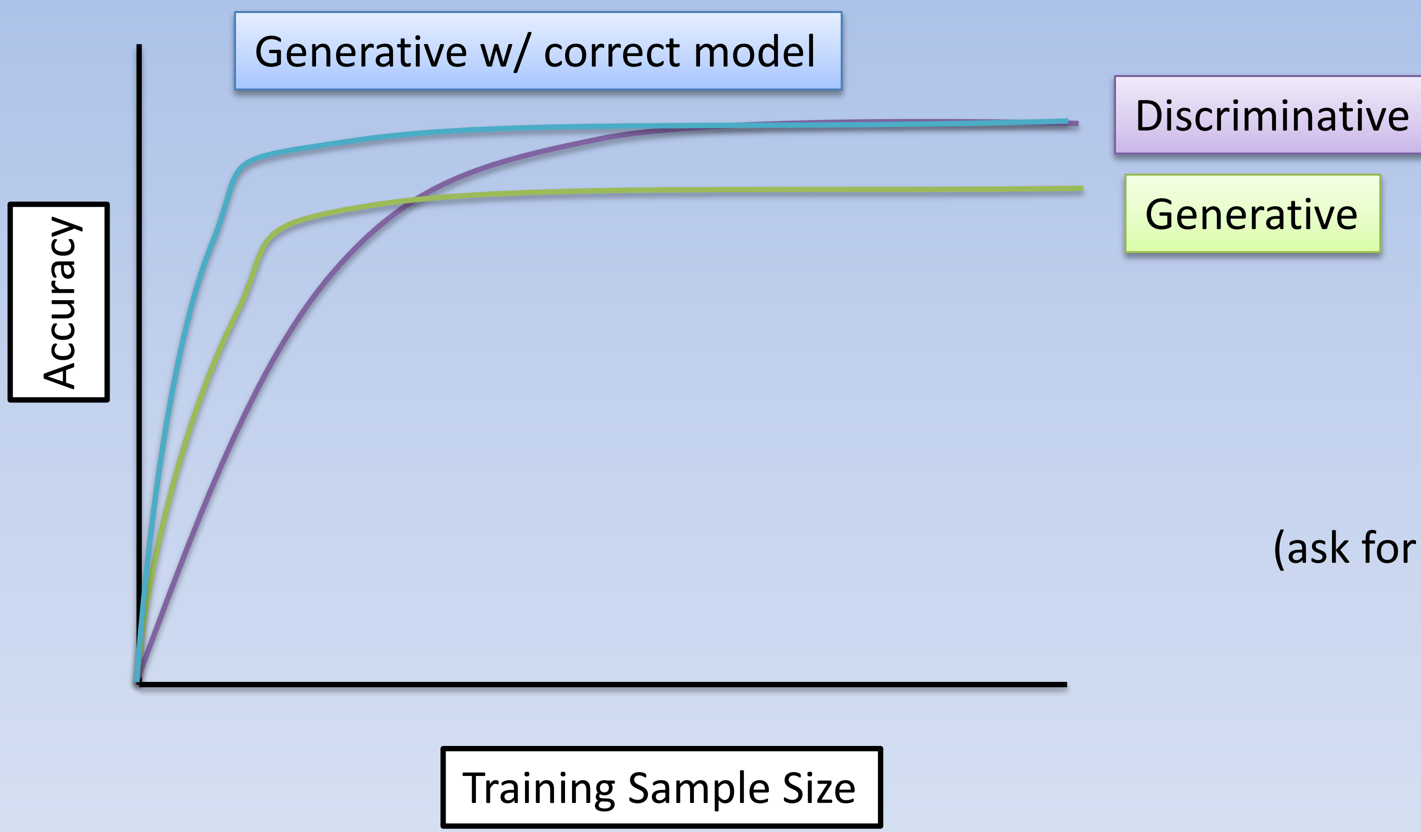
$$\text{But } \log\frac{p(Y=1\,|\,\mathbf{x})}{p(Y=-1\,|\,\mathbf{x})} = \mathbf{w}\bullet\mathbf{x} + b$$

So classify as positive iff $\mathbf{w}\bullet\mathbf{x} + b > 0$

Relationship to Naïve Bayes

- LR does not make the independence assumptions of NB
  - Can be more robust than NB, especially in the presence of irrelevant attributes
  - Also handles continuous attributes nicely
  - But (as with all discriminative models) no easy way to handle data issues such as missing data

Generative and Discriminative Pairs



(ask for paper)