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 next Thursday
  - Topics up to and including Optimization

Course Project (10/26-12/16)

- Two possibilities:
  - Propose your own. Send email by 11/2 with detailed description+group if any. Must meet with me to get approval.
  - Default: Comparative Analysis of Algorithms
- Done in groups of at most 5 people (see "Project Groups" on Canvas)
  - Can switch if allowed by both groups
- Each person in a group will read *at least two distinct* papers in the area specified for the group
  - From ICML, AAAI, NeurIPS, JMLR, MLJ, ECML etc.
  - I will send a topic and "seed" paper to each group
  - Distinct means the papers must not be tweaks of each other, exploring very similar ideas or have a significant amount of overlap
- Each person will implement *at least one* distinct algorithm and *at least one* novel extension of the algorithm
  - The minimum for a passing/"C" grade on the project
- The group will do a comparative evaluation of the algorithms implemented on *at least* 2 datasets

Course Project (10/26-12/16)

- Evaluation: 25% of grade
- Each team will make a repository on cwru-courses where you will commit the project code and store papers you read
  - csds440project-f23-*n* (*n* is your project group number. Do NOT modify/capitalize differently. Do add TAs and me as admin.)
- You may use external libraries such as pytorch for implementation
  - Again, rule is you must implement significant elements of the project on your own
- Writeup to be submitted via github (Markdown)
- Writeup/final commit is due Dec 10, 11:59pm
  - No extensions

Structure of report

- The written report will contain :
  - Individual reports with:
    - a survey of the area synthesizing the papers read
    - a description of the specific algorithms implemented, extensions and experiments
    - an insightful discussion of the results
  - A group report documenting the comparative experiments, results and discussion
- More details to follow in canvas announcement

Grading Criteria

- Thoroughness of survey
  - Did you touch on many different important ideas? How in-depth was your exploration of the ideas?
- Technical strength of implementation
  - How nontrivial were the algorithms implemented? How nontrivial was the research extension?
- Insightfulness of results and discussion
  - Beyond "A is better than B". When does a method work? Why does it work? What did your research extension do? What subsequent analysis did it inspire?
- Clarity and interestingness of writeup
  - Did you explain the ideas clearly? Did you come up with good ways to synthesize the material into coherent whole?

Grading criteria

- Each person will receive a score on each criterion
- The group as a whole will also receive a score on each criterion
- Your final grade will be 80% of your average score over all criteria + 20% of the group's average score over all criteria
- To get a more than C grade on the "Technical Strength" "Thoroughness" and "Insightfulness" criteria you will need to go beyond the minimums
  - read more papers, implement more algorithms, research multiple extensions, evaluate on more datasets, do an insightful comparison

Course Project steps

1. Collect papers (>=2 each), store in github papers/ subdirectory. Collect at least 2 datasets. Discuss as a group (or with me) to ensure everything looks reasonable (by 11/9)
2. Read and discuss papers. (by 11/16)
3. Implement algorithm(s). (by 11/30)
4. Carry out detailed comparative evaluation. Investigate parameter settings. Perform hypothesis tests.
5. Write report with your findings.

Recap

- One way to control overfitting is to use d___. Here a r___ s_____ of the nodes is left out during b_____.
- It is useful to s___ the inputs to an ANN. When done at internal nodes this is called b___ n_____.
- Nominal features have to be encoded via _____ or ___ when input to an ANN.
- Probabilistic classifiers are useful to determine the optimal hypothesis using B_____ d___ t_____.
- They also incorporate p_____ k_____ and produce c______ estimates.
- They can be g______ or d______. The first models ______, the second _______.

Today

- Probabilistic Machine Learning

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)

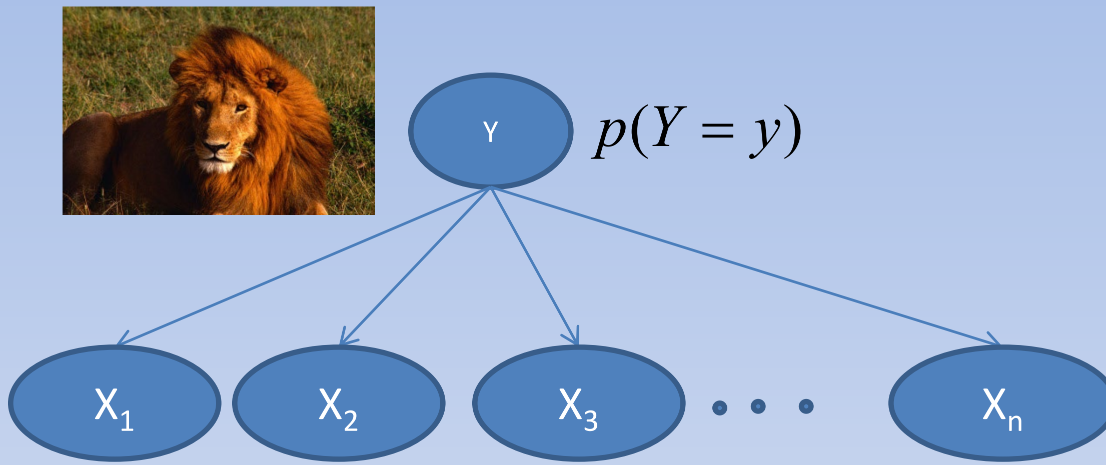
How many parameters?

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

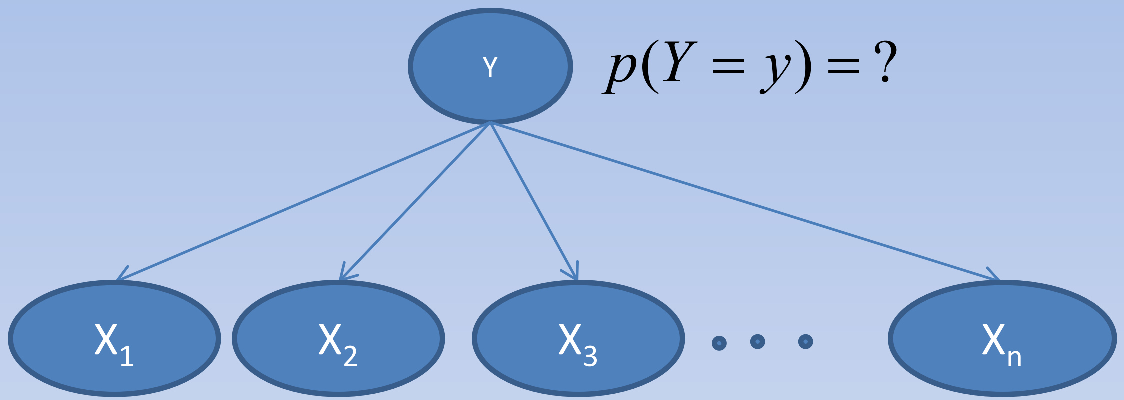
Learning a Naïve Bayes classifier

- 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:

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

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

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

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

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

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 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 $X_i=1$