Probability and naïve Bayes Classifier

Louis Oliphant
oliphant@cs.wisc.edu
cs540 section 2
Fall 2005

Announcements

- Homework 4 due Thursday
- Project
  - meet with me during office hours this week.
  - or setup a time via email
- Read
  - chapter 13
  - chapter 20 section 2 portion on Naive Bayes model (page 718)

Probability and Uncertainty

- Probability provides a way of summarizing the uncertainty that comes from our *laziness* and *ignorance*.
  - 60% chance of rain today
  - 85% chance of making a free throw
- Calculated based upon past performance, or degree of belief

Probability Notation

- Random Variables (RV):
  - are capitalized (usually) e.g. *Sky*, *RoadCurvature*, *Temperature*
  - refer to attributes of the world whose "status" is unknown
  - have one and only one value at a time
  - have a domain of values that are possible states of the world:
    - boolean: domain = <*true*, *false*>
      Cavity=true abbreviated as *cavity*
      Cavity=false abbreviated as ¬*cavity*
    - discrete: domain is countable (includes boolean)
      values are ***exhaustive and mutually exclusive***
      e.g. *Sky* domain = <*clear*, *partly_cloudy*, *overcast*>
      Sky=clear abbreviated as *clear*
      Sky≠clear also abbrv. as ¬*clear*
    - continuous:domain is real numbers (beyond scope of CS540)

Probability Notation

- An agent's uncertainty is represented by:

P(*A*=*a*) or simply *P*(*a*), this is:

  - the agent's degree of belief that variable *A* takes on value *a* given no other information relating to *A*
  - a single probability called an unconditional or prior probability

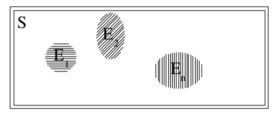
- Properties of P(A=a):
  - $0 \le P(a) \le 1$
  - $\Sigma P(a_i) = P(a_1) + P(a_2) + ... + P(a_n) = 1$

    sum over all values in the domain of variable *A* is 1
    because domain is exhaustive and mutually exclusive

Axioms of Probability

- S – Sample Space (set of possible outcomes)
- E – Some Event (some subset of outcomes)
- Axioms:
  - $0 \le P(E) \le 1$
  - P(S)=1
  - for any sequence of mutually exclusive events, $E_1, E_2, ...E_n$

    $P(E_1 \text{ or } E_2 ... E_n) = P(E_1)+P(E_2)+...+P(E_n)$



Probability Table

Outlook		
sunny	overcast	rainy
5/14	4/14	5/14

- P(Weather=sunny)=P(sunny)=5/13
- P(Weather)={5/14, 4/14, 5/14}

- Calculate probabilities from data

A Hypothesis for the Circus

Day	Outlook	Temperature	Humidity	Wind	>1,000?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Joint Probability Table

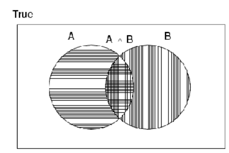
		Outlook		
		sunny	overcast	rainy
Temperature	hot	2/14	2/14	0/14
	mild	2/14	1/14	3/14
	cool	1/14	1/14	2/14

P(Outlook=sunny, Temperature=hot) = P(sunny,hot) = 2/14
P(Temperature=hot) = P(hot) = 2/14+2/14+0/14 = 4/14

With N Random variables that can take k values the full joint probability table size is k^N

Probability of Disjunctions

- P(A or B) = P(A) + P(B) – P(A and B)
- P(Outlook=sunny or Temperature=hot)?
  - P(sunny) + P(hot) – P(sunny,hot)
  - 5/14 + 4/14 - 2/14



Marginalization

- P(cavity)=0.108+0.012+0.072+0.008=0.2
- Called summing out or marginalization

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

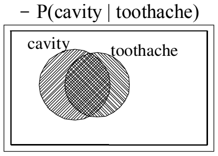
Conditional Probability

- Probabilities discussed up until now are called prior probabilities or unconditional probabilities
  - Probabilities depend only on the data, not on any other variable
- But what if you have some evidence or knowledge about the situation? You know you have a toothache. Now what is the probability of having a cavity?

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	**.108**	**.012**	**.072**	**.008**
¬ cavity	**.016**	**.064**	**.144**	**.576**

Conditional Probability

- Written like P(A | B)
  - P(cavity | toothache)



	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

Calculate conditional probabilities from data as follows:
P(A | B) = P(A,B) / P(B) if P(B)≠0
P(cavity | toothache) = (0.108 + 0.012) / (0.108 + 0.012 + 0.016 + 0.064)
P(cavity | toothache) = 0.12 / 0.2 = 0.6
What is P(no cavity | toothache) ?

Conditional Probability

- P(A | B) = P(A,B) / P(B)
- You can think of P(B) as just a normalization constant to make P(A|B) adds up to 1.

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- Product rule: P(A,B) = P(A|B)P(B) = P(B|A)P(A)
- Chain Rule is successive applications of product rule:

$$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_1,\ldots,X_{n-1})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$$
$$= \mathbf{P}(X_1,\ldots,X_{n-2})\, \mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\, \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$$
$$= \ldots$$
$$= \boxed{\prod_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})}$$

Independence

- What if I know Weather=cloudy today. Now what is the P(cavity)?
- if knowing some evidence doesn't change the probability of some other random variable then we say the two random variables are independent
- A and B are independent if P(A|B)=P(A).
- Other ways of seeing this (all are equivalent):
  - P(A|B)=P(A)
  - P(A,B)=P(A)P(B)
  - P(B|A)=P(B)
- Absolute Independence is powerful but rare!

Conditional Independence

- **P**(*Toothache, Cavity, Catch*) has $2^3 - 1 = 7$ independent entries
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  (1) **P**(*catch* | *toothache, cavity*) = **P**(*catch* | *cavity*)
- The same independence holds if I haven't got a cavity:
  (2) **P**(*catch* | *toothache,¬cavity*) = **P**(*catch* | ¬*cavity*)
- *Catch* is conditionally independent of *Toothache* given *Cavity*:
  P(*Catch* | *Toothache,Cavity*) = **P**(*Catch* | *Cavity*)
- Equivalent statements:
  P(*Toothache* | *Catch, Cavity*) = **P**(*Toothache* | *Cavity*)
  P(*Toothache, Catch* | *Cavity*) = **P**(*Toothache* | *Cavity*) **P**(*Catch* | *Cavity*)

Bayes' Rule

- Remember Conditional Probabilities:
  - P(A|B)=P(A,B)/P(B)
  - P(B)P(A|B)=P(A,B)

  - P(B|A)=P(B,A)/P(A)
  - P(A)P(B|A)=P(B,A)

  - P(B,A)=P(A,B)

  - P(B)P(A|B)=P(A)P(B|A)
  - Bayes' Rule: P(A|B)=P(B|A)P(A) / P(B)

Bayes' Rule

- P(A|B)=P(B|A)P(A) / P(B)
- A more general form is:
  - P(Y|X,e)=P(X|Y,e)P(Y|e) / P(X|e)
- Bayes' rule allows you to turn conditional probabilities on their head:
  - Useful for assessing diagnostic probability from causal probability:
  - P(Cause|Effect) = P(Effect|Cause) P(Cause) / P(Effect)
  - E.g., let *M* be meningitis, *S* be stiff neck:
    P(m|s) = P(s|m) P(m) / P(s) = 0.8 × 0.0001 / 0.1 = 0.0008
  - Note: posterior probability of meningitis still very small!

Bayes' Rule used in Classification

Day	Outlook	Temperature	Humidity	Wind	>1,000?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Overcast	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

What is the probability of >1,000 people given Outlook=overcast, Temperature=mild, Humidity=normal, Wind=weak?
Use Bayes' Rule and Assume Features are independent given the class

naïve Bayes (Idiot's Bayes) model

$\mathbf{P}(\text{Class}|\text{Feature}_1, \ldots ,\text{Feature}_n) = \mathbf{P}(\text{Class}) \Pi_i \mathbf{P}(\text{Feature}_i|\text{Class})$
classify with highest probability

- One of the most widely used classifiers
- Very Fast to train and to classify
  - One pass over all data to train
  - One lookup for each feature / class combination to classify
- Assuming the features are independent given the class (conditional independence)

Issues with naïve Bayes

- In practice, we estimate the probabilities by maintaining counts as we pass through the training data, and then divide through at the end
- But what happens if, when classifying, we come across a feature / class combination that wasn't see in training?

$$P(x_n|c)=0 \quad \ldots \textit{ therefore} \ldots$$
$$P(c)\times\prod_n P(x_n|c)=0$$

•Typically, we can get around this by initializing all the counts to Laplacian priors (small uniform values, e.g., 1) instead of 0

- This way, the probability will still be small, but not impossible
- This is also called "smoothing"

Issues with naïve Bayes

- Another big problem with naïve Bayes: often the conditional independence assumption is violated
  - Consider the task of classifying whether or not a certain word is a corporation name
    - e.g. "Google," "Microsoft," "IBM," and "ACME"
  - Two useful features we might want to use are `captialized`, and `all-capitals`
  - Naïve Bayes will assume that these two features are independent given the class, but this clearly isn't the case (things that are all-caps must also be capitalized)!!
- However naïve Bayes seems to work well in practice even when this assumption is violated

Training with naïve Bayes

Day	Outlook	Temperature	Humidity	Wind	>1,000?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

- Use table to calculate probabilities
  - table for class, and for each feature / class combination

Conclusion

- Probabilities
- Joint Probabilities
- Conditional Probabilities
- Independence, Conditional Independence
- naïve Bayes Classifier