Bayes optimal classifier
Naïve Bayes

Machine Learning – CSE446

Carlos Guestrin

University of Washington

November 18, 2014

Classification

- **Learn**: h:**X** $\mapsto$ Y
  - **X** – features
  - Y – target classes

- Suppose you know P(Y|**X**) exactly, how should you classify?
  - Bayes optimal classifier:

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$(\forall i,j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}$$

How hard is it to learn the optimal classifier?

- Data =

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

- How do we represent these? How many parameters?
  - Prior, P(Y):
    - Suppose Y is composed of *k* classes
  - Likelihood, P(**X**|Y):
    - Suppose **X** is composed of *d* binary features
- Complex model ! High variance with limited data!!!

Conditional Independence

- X is **conditionally independent** of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = i | Y = j, Z = k) = P(X = i | Z = k)$$

- e.g., $P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$

- Equivalent to:

$$P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$$

What if features are independent?

- Predict Thunder
- From two **conditionally Independent** features
  - Lightening
  - Rain

The Naïve Bayes assumption

- Naïve Bayes assumption:
  - Features are independent given class:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

  - More generally:

$$P(X_1 \ldots X_d|Y) = \prod_i P(X_i|Y)$$

- How many parameters now?
  - Suppose **X** is composed of *d* binary features

The Naïve Bayes Classifier

- Given:
  - Prior P(Y)
  - *d* conditionally independent features **X** given the class Y
  - For each X_i, we have likelihood $P(X_i|Y)$

- Decision rule:

$$y^* = h_{NB}(\mathbf{x}) = \arg\max_y P(y)P(x_1, \ldots, x_d \mid y)$$
$$= \arg\max_y P(y) \prod_i P(x_i|y)$$

- If assumption holds, NB is optimal classifier!

MLE for the parameters of NB

- Given dataset
  - Count(A=a,B=b) == number of examples where A=a and B=b
- MLE for NB, simply:
  - Prior: P(Y=y) =
  - Likelihood: $P(X_i=x_i|Y=y)$ =

Subtleties of NB classifier 1 – Violating the NB assumption

- Usually, features are not conditionally independent:

$$P(X_1 \ldots X_d|Y) \neq \prod_i P(X_i|Y)$$

- Actual probabilities P(Y|**X**) often biased towards 0 or 1
- Nonetheless, NB is the single most used classifier out there
  - NB often performs well, even when assumption is violated
  - [Domingos & Pazzani '96] discuss some conditions for good performance

Subtleties of NB classifier 2 – Insufficient training data

- What if you never see a training instance where X_1=a when Y=b?
  - e.g., Y={SpamEmail}, X_1={'Enlargement'}
  - $P(X_1=a \mid Y=b) = 0$
- Thus, no matter what the values $X_2,\ldots,X_d$ take:
  - $P(Y=b \mid X_1=a,X_2,\ldots,X_d) = 0$
- "Solution": smoothing
  - Add "fake" counts, usually uniformly distributed
  - Equivalent to Bayesian Learning

Text classification

- Classify e-mails
  - Y = {Spam,NotSpam}
- Classify news articles
  - Y = {what is the topic of the article?}
- Classify webpages
  - Y = {Student, professor, project, …}
- What about the features **X**?
  - The text!

Features **X** are entire document – X_i for i^{th} word in article

Article from rec.sport.hockey

```
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most
obvious candidate for pleasant surprise is Alex
Zhitnik. He came highly touted as a defensive
defenseman, but he's clearly much more than that.
Great skater and hard shot (though wish he were
more accurate). In fact, he pretty much allowed
the Kings to trade away that huge defensive
liability Paul Coffey. Kelly Hrudey is only the
biggest disappointment if you thought he was any
good to begin with. But, at best, he's only a
mediocre goaltender. A better choice would be
Tomas Sandstrom, though not through any fault of
his own, but because some thugs in Toronto decided
```


NB for Text classification

- P(**X**|Y) is huge!!!
  - Article at least 1000 words, $\mathbf{X}=\{X_1,\ldots,X_{1000}\}$
  - X_i represents i^{th} word in document, i.e., the domain of X_i is entire vocabulary, e.g., Webster Dictionary (or more), 10,000 words, etc.
- NB assumption helps a lot!!!
  - $P(X_i=x_i|Y=y)$ is just the probability of observing word x_i in a document on topic y

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

Bag of words model

- Typical additional assumption – **Position in document doesn't matter**: $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$
  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

When the lecture is over, remember to wake up the person sitting next to you in the lecture room.

Bag of words model

- Typical additional assumption – **Position in document doesn't matter**: $P(X_i=x_i|Y=y) = P(X_k=x_i|Y=y)$
  - "Bag of words" model – order of words on the page ignored
  - Sounds really silly, but often works very well!

$$P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

in is lecture lecture next over person remember room sitting the the the to to up wake when you

Bag of Words Approach

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

NB with Bag of Words for text classification

- Learning phase:
  - Prior P(Y)
    - Count how many documents you have from each topic (+ prior)
  - $P(X_i|Y)$
    - For each topic, count how many times you saw word in documents of this topic (+ prior)
- Test phase:
  - For each document
    - Use naïve Bayes decision rule

$$h_{NB}(\mathbf{x}) = \arg\max_y P(y) \prod_{i=1}^{LengthDoc} P(x_i|y)$$

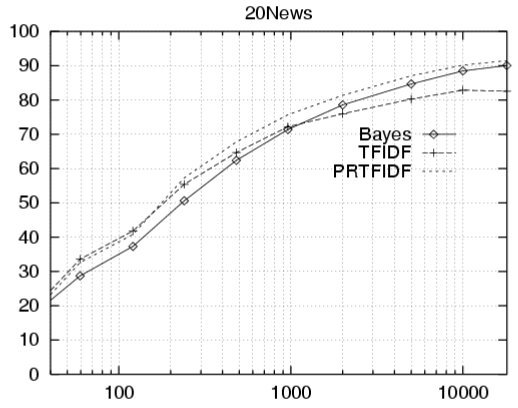
Twenty News Groups results

Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

comp.graphics	misc.forsale
comp.os.ms-windows.misc	rec.autos
comp.sys.ibm.pc.hardware	rec.motorcycles
comp.sys.mac.hardware	rec.sport.baseball
comp.windows.x	rec.sport.hockey
alt.atheism	sci.space
soc.religion.christian	sci.crypt
talk.religion.misc	sci.electronics
talk.politics.mideast	sci.med
talk.politics.misc	
talk.politics.guns	

Naive Bayes: 89% classification accuracy

Learning curve for Twenty News Groups



Accuracy vs. Training set size (1/3 withheld for test)

Bayesian Networks – Representation

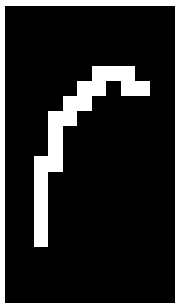
Machine Learning – CSE446

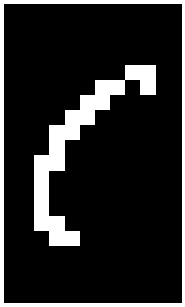
Carlos Guestrin

University of Washington

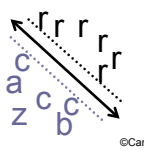
November 18, 2014

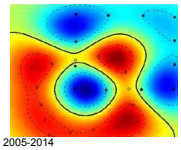
Handwriting recognition





Character recognition, e.g., kernel SVMs





Webpage classification

→ Company home page

vs

Personal home page

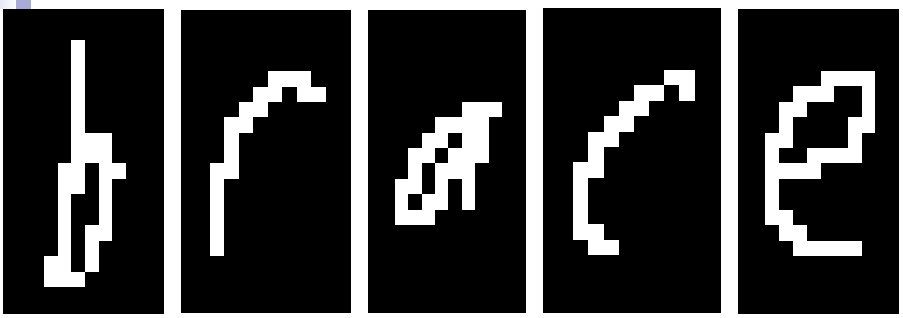
vs

University home page

vs

…

Handwriting recognition 2



Webpage classification 2

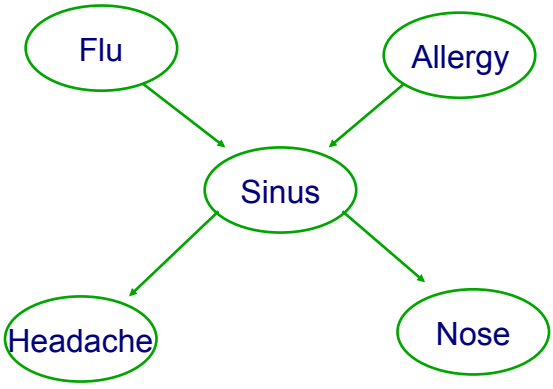
Today – Bayesian networks

- One of the most exciting advancements in statistical AI in the last decades
- Generalizes naïve Bayes and logistic regression classifiers
- Compact representation for exponentially-large probability distributions
- Exploit conditional independencies

Causal structure

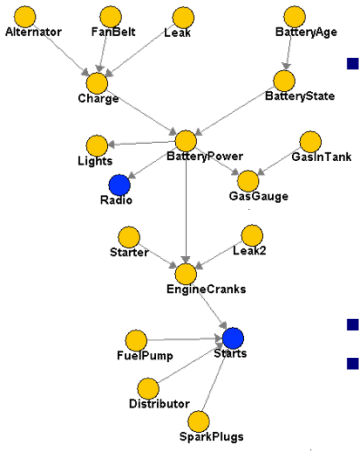
- Suppose we know the following:
  - The flu causes sinus inflammation
  - Allergies cause sinus inflammation
  - Sinus inflammation causes a runny nose
  - Sinus inflammation causes headaches
- How are these connected?

Possible queries



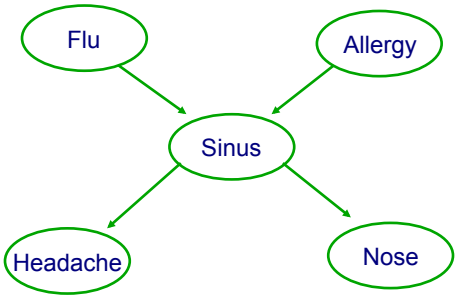
- Inference
- Most probable explanation
- Active data collection

Car starts BN

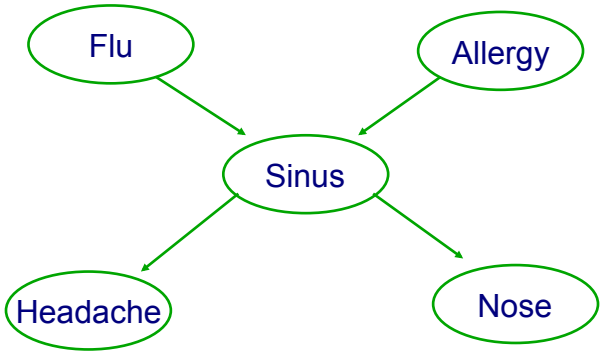


- 18 binary attributes
- Inference
  - P(BatteryAge|Starts=f)
- 2^{16} terms, why so fast?
- Not impressed?
  - HailFinder BN – more than 3^{54} = 58149737003040059690390169 terms

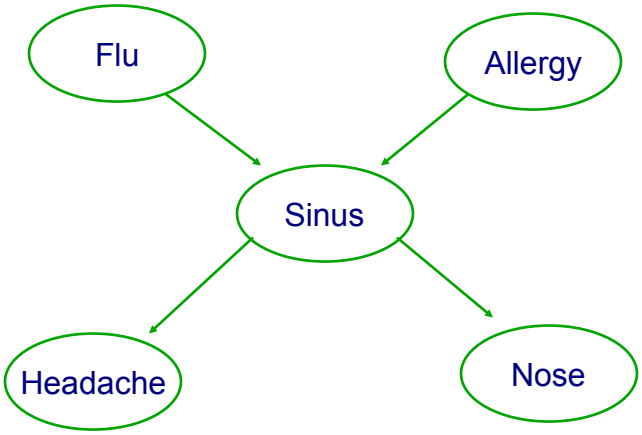
Factored joint distribution - Preview



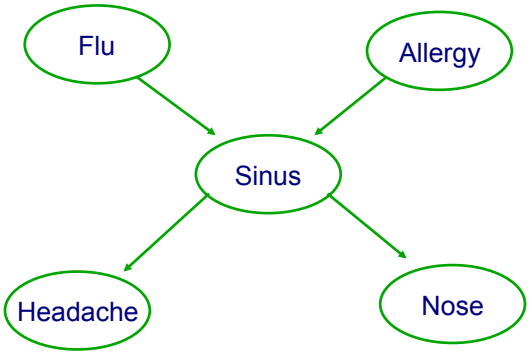
What about probabilities? Conditional probability tables (CPTs)



Number of parameters



Key: Independence assumptions



Knowing sinus separates the variables from each other

(Marginal) Independence

- Flu and Allergy are (marginally) independent

Flu = t	
Flu = f	

Allergy = t	
Allergy = f	

	Flu = t	Flu = f
Allergy = t		
Allergy = f		

Marginally independent random variables

- **Sets** of variables **X**, **Y**
- X is independent of Y if
  - $P \models (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y}), \forall \mathbf{x} \in \text{Val}(\mathbf{X}), \mathbf{y} \in \text{Val}(\mathbf{Y})$
- Shorthand:
  - **Marginal independence:** $P \models (\mathbf{X} \perp \mathbf{Y})$
- **Proposition:** *P* statisfies $(\mathbf{X} \perp \mathbf{Y})$ if and only if
  - $P(\mathbf{X},\mathbf{Y}) = P(\mathbf{X})\, P(\mathbf{Y})$

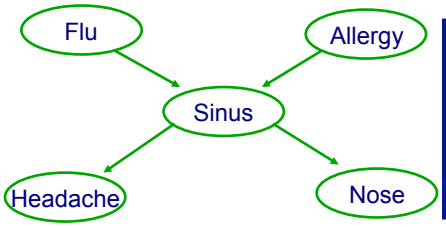
Conditional independence

- Flu and Headache are not (marginally) independent
- Flu and Headache are independent given Sinus infection
- More Generally:

Conditionally independent random variables

- **Sets** of variables **X**, **Y**, **Z**
- X is independent of Y given Z if
  - $P \models (\mathbf{X}=\mathbf{x} \perp \mathbf{Y}=\mathbf{y} | \mathbf{Z}=\mathbf{z}), \forall \mathbf{x} \in Val(\mathbf{X}), \mathbf{y} \in Val(\mathbf{Y}), \mathbf{z} \in Val(\mathbf{Z})$
- Shorthand:
  - **Conditional independence:** $P \models (\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$
  - For $P \models (\mathbf{X} \perp \mathbf{Y} | \varnothing)$, write $P \models (\mathbf{X} \perp \mathbf{Y})$
- **Proposition:** *P* statisfies $(\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z})$ if and only if
  - $P(\mathbf{X},\mathbf{Y}|\mathbf{Z}) = P(\mathbf{X}|\mathbf{Z})\, P(\mathbf{Y}|\mathbf{Z})$

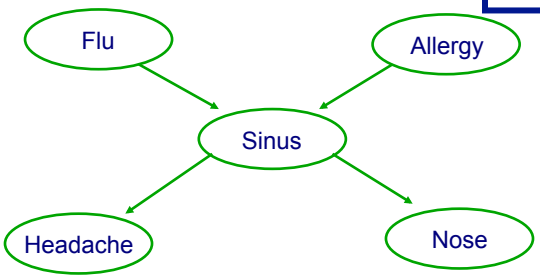
The independence assumption



Local Markov Assumption: A variable X is independent of its non-descendants given its parents

Explaining away

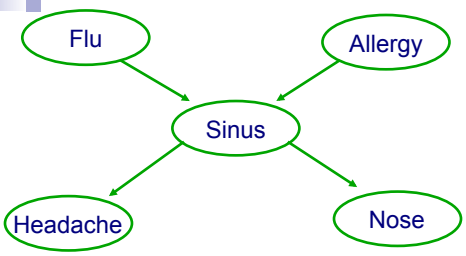
Local Markov Assumption:
A variable X is independent of its non-descendants given its parents



Naïve Bayes revisited

Local Markov Assumption:
A variable X is independent of its non-descendants given its parents

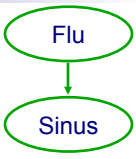
Joint distribution



Why can we decompose? Markov Assumption!

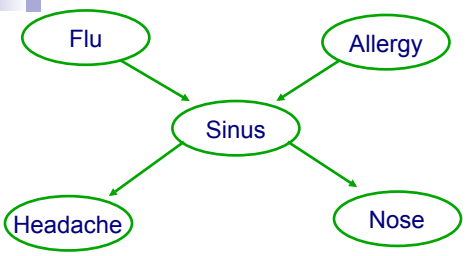
The chain rule of probabilities

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



- More generally:
  - $P(X_1,\ldots,X_n) = P(X_1)\, P(X_2|X_1) \ \ldots \ P(X_n|X_1,\ldots,X_{n-1})$

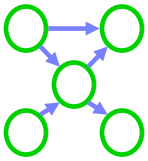
Chain rule & Joint distribution



Local Markov Assumption:
A variable X is independent of its non-descendants given its parents

The Representation Theorem – Joint Distribution to BN

BN:



Encodes independence assumptions

If conditional independencies in BN are subset of conditional independencies in *P*

Obtain

Joint probability distribution:

$$P(X_1,\ldots,X_n) = \prod_{i=1}^{n} P\left(X_i \mid \mathbf{Pa}_{X_i}\right)$$

Two (trivial) special cases

Edgeless graph

Fully-connected graph

Feature selection

- Want to learn f:**X**→Y
  - **X**=<X_1,…,X_n>
  - but some features are more important than others
- **Approach**: select subset of features to be used by learning algorithm
  - **Score** each feature (or sets of features)
  - **Select** set of features with best score

Simple greedy **forward** feature selection algorithm

- Pick a dictionary of features
  - e.g., polynomials for linear regression
- Greedy heuristic:
  - Start from empty (or simple) set of features $F_0 = \emptyset$
  - Run learning algorithm for current set of features F_t
    - Obtain h_t
  - Select **next best feature X_i**
    - e.g., X_j that results in lowest cross-validation error learner when learning with $F_t \cup \{X_j\}$
  - $F_{t+1} \leftarrow F_t \cup \{X_i\}$
  - Recurse

Simple greedy **backward** feature selection algorithm

- Pick a dictionary of features
  - e.g., polynomials for linear regression
- Greedy heuristic:
  - Start from all features $F_0 = F$
  - Run learning algorithm for current set of features F_t
    - Obtain h_t
  - Select **next worst feature X_i**
    - e.g., X_j that results in lowest cross-validation error learner when learning with $F_t - \{X_j\}$
  - $F_{t+1} \leftarrow F_t - \{X_i\}$
  - Recurse

Impact of feature selection on classification of fMRI data [Pereira et al. '05]

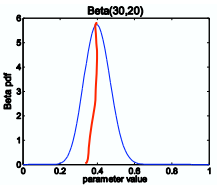
Accuracy classifying category of word read by subject

#voxels	mean	subjects							
		233B	329B	332B	424B	474B	496B	77B	86B
50	0.735	0.783	0.817	0.55	0.783	0.75	0.8	0.65	0.75
100	0.742	0.767	0.8	0.533	0.817	0.85	0.783	0.6	0.783
200	0.737	0.783	0.783	0.517	0.817	0.883	0.75	0.583	0.783
300	**0.75**	**0.8**	**0.817**	**0.567**	**0.833**	**0.883**	**0.75**	**0.583**	**0.767**
400	0.742	0.8	0.783	0.583	0.85	0.833	0.75	0.583	0.75
800	0.735	0.833	0.817	0.567	0.833	0.833	0.7	0.55	0.75
1600	0.698	0.8	0.817	0.45	0.783	0.833	0.633	0.5	0.75
all (~2500)	0.638	0.767	0.767	0.25	0.75	0.833	0.567	0.433	0.733

Table 1: **Average accuracy across all pairs of categories, restricting the procedure to use a certain number of voxels for each subject.** The highlighted line corresponds to the best mean accuracy, obtained using 300 voxels.

Voxels scored by p-value of regression to predict voxel value from the task

MAP for Beta distribution



$$P(\theta \mid \mathcal{D}) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H+\alpha_H, \beta_T+\alpha_T)} \sim Beta(\beta_H+\alpha_H, \beta_T+\alpha_T)$$

- MAP: use most likely parameter:

$$\widehat{\theta} = \arg\max_{\theta} P(\theta \mid \mathcal{D}) = \frac{\beta_H+\alpha_H-1}{\beta_H+\alpha_H+\beta_T+\alpha_T-2}$$

- Beta prior equivalent to extra thumbtack flips
- As $N \rightarrow 1$, prior is "forgotten"
- **But, for small sample size, prior is important!**

Bayesian learning for NB parameters – a.k.a. smoothing

- Dataset of *N* examples
- Prior
  - "distribution" $Q(X_i,Y)$, $Q(Y)$
  - *m* "virtual" examples
- MAP estimate
  - $P(X_i|Y)$
- Now, even if you never observe a feature/class, posterior probability never zero

Properties of independence

- **Symmetry:**
  - (**X** ⊥ **Y** | **Z**) ⇒ (**Y** ⊥ **X** | **Z**)
- **Decomposition:**
  - (**X** ⊥ **Y**,**W** | **Z**) ⇒ (**X** ⊥ **Y** | **Z**)
- **Weak union:**
  - (**X** ⊥ **Y**,**W** | **Z**) ⇒ (**X** ⊥ **Y** | **Z**,**W**)
- **Contraction:**
  - (**X** ⊥ **W** | **Y**,**Z**) & (**X** ⊥ **Y** | **Z**) ⇒ (**X** ⊥ **Y**,**W** | **Z**)
- **Intersection:**
  - (**X** ⊥ **Y** | **W**,**Z**) & (**X** ⊥ **W** | **Y**,**Z**) ⇒ (**X** ⊥ **Y**,**W** | **Z**)
  - Only for positive distributions!
  - $P(\alpha)>0$, $\forall\alpha$, $\alpha>0$;