CMSC 471:
Probability, and Reasoning and Learning with Uncertainty (Bayesian Reasoning)

Chapters 12 & 13

KMA Solaiman – ksolaima@umbc.edu

Some slides courtesy Tim Finin and Frank Ferraro

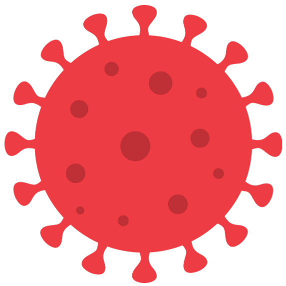
Today's topics

- Motivation
- Review probability theory
- Bayesian inference
  - From the joint distribution
  - Using independence/factoring
  - From sources of evidence
- Naïve Bayes algorithm for inference and classification tasks

Motivation: causal reasoning

- As the sun rises, the rooster crows
  - Does this correlation imply causality?
  - If so, which way does it go?
- The evidence can come from
  - Probabilities and Bayesian reasoning
  - Common sense knowledge
  - Experiments
- Bayesian Belief Networks (BBNs) are useful for modeling causal reasoning

Motivation: logic isn’t enough



- Classical logic is designed to work with certainties
- Getting a positive result on a COVID test doesn’t necessarily mean you are infected
- And a negative result doesn’t necessarily mean you are not infected
- You need to know the **true/false positive** and **true/false negative** rates of the test

Decision making with uncertainty 🤔

Rational behavior: for each possible action:

- Identify possible outcomes and **for each**
  - Compute **probability** of outcome
  - Compute **utility** of outcome
  - Compute probability-weighted **(expected) utility** of outcome
- Select action with the highest expected utility (principle of **Maximum Expected Utility**)

Consider

- Your house has an alarm system
- It should go off if a burglar breaks into the house
- It can also go off if there is an earthquake
- How can we predict what's happened if the alarm goes off?
  - Someone has broken in!
  - It's a minor earthquake

Probability theory 101

- **Random variables:**
  - Domain
- **Atomic event**: complete specification of state
- **Prior probability**: degree of belief without any other evidence or info
- **Joint probability**: matrix of combined probabilities of set of variables

- Alarm, Burglary, Earthquake

Boolean (these) or discrete (0-9), continuous (float)

- Alarm=T∧Burglary=T∧Earthquake=F
  alarm ∧ burglary ∧ ¬earthquake

- P(Burglary) = 0.1
  P(Alarm) = 0.1
  P(earthquake) = 0.000003

- P(Alarm, Burglary) =

	alarm	**¬alarm**
burglary	.09	.01
¬burglary	.1	.8

Probability theory 101

	alarm	¬alarm
burglary	.09	.01
¬burglary	.1	.8

- **Conditional probability**: prob. of effect given causes
- **Computing conditional probs**:
  - $P(a \mid b) = P(a \wedge b) / P(b)$
  - $P(b)$: **normalizing** constant
- **Product rule**:
  - $P(a \wedge b) = P(a \mid b) * P(b)$
- **Marginalizing**:
  - $P(B) = \Sigma_a P(B, a)$
  - $P(B) = \Sigma_a P(B \mid a) P(a)$ (**conditioning**)

- P(burglary | alarm) = .47
  P(alarm | burglary) = .9
- P(burglary | alarm) =
  P(burglary ∧ alarm) / P(alarm)
  = .09/.19 = .47
- P(burglary ∧ alarm) =
  P(burglary | alarm) * P(alarm)
  = .47 * .19 = .09
- P(alarm) =
  P(alarm ∧ burglary) +
  P(alarm ∧ ¬burglary)
  = .09+.1 = .19

Probability theory 101

	alarm	¬alarm
burglary	.09	.01
¬burglary	.1	.8

- **Conditional probability**: prob. of effect given causes
- **Computing conditional probs**:
  - $P(a \mid b) = P(a \wedge b) / P(b)$
  - $P(b)$: **normalizing** constant
- **Product rule**:
  - $P(a \wedge b) = P(a \mid b) * P(b)$
- **Marginalizing**:
  - $P(B) = \Sigma_a P(B, a)$
  - $P(B) = \Sigma_a P(B \mid a) P(a)$ (**conditioning**)

- $P(\text{burglary} \mid \text{alarm}) = .47$
  $P(\text{alarm} \mid \text{burglary}) = .9$
- $P(\text{burglary} \mid \text{alarm}) =$
  $P(\text{burglary} \wedge \text{alarm}) / P(\text{alarm})$
  $= .09/.19 = .47$
- $P(\text{burglary} \wedge \text{alarm}) =$
  $P(\text{burglary} \mid \text{alarm}) * P(\text{alarm})$
  $= .47 * .19 = .09$
- $P(\text{alarm}) =$
  $P(\text{alarm} \wedge \text{burglary}) +$
  $P(\text{alarm} \wedge \neg\text{burglary})$
  $= .09+.1 = .19$

Example: Inference from the joint

	alarm		¬alarm	
	earthquake	¬earthquake	earthquake	¬earthquake
burglary	.01	.08	.001	.009
¬burglary	.01	.09	.01	.79

P(burglary | alarm) = α P(burglary, alarm)

= α [P(burglary, alarm, earthquake) + P(burglary, alarm, ¬earthquake)

= α [(.01, .01) + (.08, .09)]

= α [(.09, .1)]

Since P(burglary | alarm) + P(¬burglary | alarm) = 1, α = 1/(.09+.1) = 5.26

(i.e., P(alarm) = 1/α = .19 – **quizlet**: how can you verify this?)

P(burglary | alarm) = .09 * 5.26 = .474

P(¬burglary | alarm) = .1 * 5.26 = .526

Consider

- A student has to take an exam
  - She might **be smart**
  - She might **have studied**
  - She may **be prepared** for the exam
- How are these related?
- We can collect joint probabilities for the three events
  - Measure “prepared” as “got a passing grade”

Exercise: Inference from the joint

p(smart ∧ study ∧ prepared)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Each of the 8 highlighted boxes has the joint probability for the three values of smart, study, prepared

Standard way to show joint probabilities of 3 variables as a 2D table

Queries:

- **What is the prior probability of *smart*?**
- What is the prior probability of *study*?
- What is the conditional probability of *prepared*, given *study* and *smart*?

Exercise:
Inference from the joint

p(smart ∧ study ∧ prepared)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

- **What is the prior probability of *smart*?**
- What is the prior probability of *study*?
- What is the conditional probability of *prepared*, given *study* and *smart*?

p(smart) = .432 + .16 + .048 + .16 = **0.8**

Exercise:
Inference from the joint

p(smart ∧ study ∧ prepared)	smart		¬smart	
	study	**¬study**	**study**	**¬study**
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

- What is the prior probability of *smart*?
- **What is the prior probability of *study*?**
- What is the conditional probability of *prepared*, given *study* and *smart*?

Exercise:
Inference from the joint

p(smart ∧ study ∧ prepared)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

- What is the prior probability of *smart*?
- **What is the prior probability of *study*?**
- What is the conditional probability of *prepared*, given *study* and *smart*?

p(study) = .432 + .048 + .084 + .036 = **0.6**

Exercise:
Inference from the joint

p(smart ∧ study ∧ prepared)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

- What is the prior probability of *smart*?
- What is the prior probability of *study*?
- **What is the conditional probability of *prepared*, given *study* and *smart*?**

Exercise:
Inference from the joint

p(smart ∧ study ∧ prepared)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

– What is the prior probability of *smart*?

– What is the prior probability of *study*?

– **What is the conditional probability of *prepared*, given *study* and *smart*?**

p(prepared|smart,study)= p(prepared,smart,study)/p(smart, study)
= .432 / (.432 + .048)
= **0.9**

Independence

- When variables don’t affect each others’ probabilities, they are **independent;** we can easily compute their joint & conditional probability:

  Independent(A, B) → $P(A \wedge B) = P(A) * P(B)$; $P(A|B) = P(A)$

- {moonPhase, lightLevel} *might* be independent of {burglary, alarm, earthquake}
  - Maybe not: burglars may be more active during a new moon because darkness hides their activity
  - But if we know light level, moon phase doesn’t affect whether we are burglarized
  - If burglarized, light level doesn’t affect if alarm goes off
- Need a more complex notion of independence and methods for **reasoning about the relationships**

Exercise: Independence

p(smart ∧ study ∧ prepared)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Queries:

- **Q1: Is *smart* independent of *study*?**
- **Q2: Is *prepared* independent of *study*?**

How can we tell?

Exercise: Independence

p(smart ∧ study ∧ prepared)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Q1: Is *smart* independent of *study*?

- You might have some **intuitive beliefs** based on your experience
- You can also **check the data**

Which way to answer this is better?

Exercise: Independence

p(smart ∧ study ∧ prepared)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Q1: Is *smart* independent of *study*?

Q1 true iff p(smart|study) == p(smart)

p(smart) = .432 + 0.048 + .16 + .16 = **0.8**

p(smart|study) = p(smart,study)/p(study)

= (.432 + .048) / .6 = 0.48/.6 = **0.8**

0.8 == 0.8 ∴ **smart is independent of study**

Exercise: Independence

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Q2: Is *prepared* independent of *study*?

- What is prepared?
- Q2 true iff

Exercise: Independence

p(smart ∧ study ∧ prep)	smart		¬smart	
	study	¬study	study	¬study
prepared	.432	.16	.084	.008
¬prepared	.048	.16	.036	.072

Q2: Is *prepared* independent of *study*?

Q2 true iff p(prepared|study) == p(prepared)

p(prepared) = .432 + .16 + .84 + .008 = .684

p(prepared|study) = p(prepared,study)/p(study)

= (.432 + .084) / .6 = .86

0.86 ≠ 0.684, ∴ **prepared not independent of study**

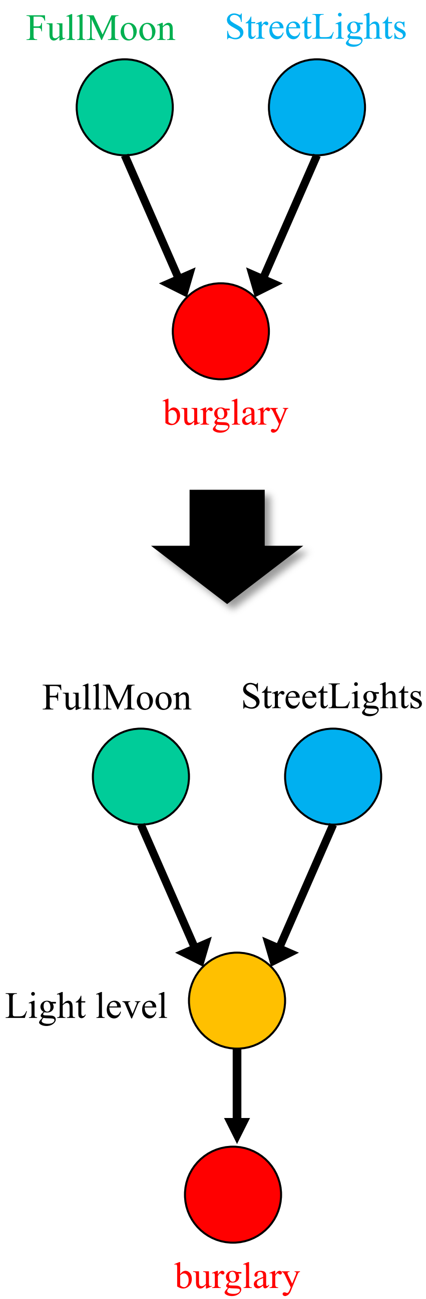
Absolute & conditional independence

- Absolute independence:
  - A and B are **independent** if $P(A \wedge B) = P(A) * P(B)$; equivalently, $P(A) = P(A \mid B)$ and $P(B) = P(B \mid A)$
- A and B are **conditionally independent** given C if
  - $P(A \wedge B \mid C) = P(A \mid C) * P(B \mid C)$

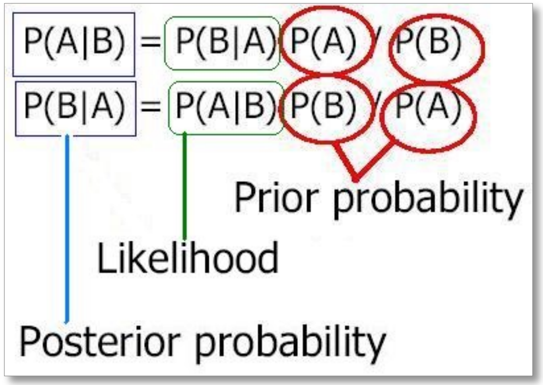
  If it holds, lets us decompose the joint distribution:
  - $P(A \wedge B \wedge C) = P(A \mid C) * P(B \mid C) * P(C)$
- Moon-Phase and Burglary are ***conditionally independent given*** Light-Level
- Conditional independence is weaker than absolute independence, but useful in decomposing full joint probability distribution

Conditional independence

- Conditional independence often comes from **causal relations**
  - FullMoon causally affects LightLevel at night as does StreetLights
- In burglary scenario, FullMoon doesn't affect anything else
- Knowing *LightLevel*, we can ignore *FullMoon* and *StreetLights* *when* predicting if alarm suggests Burglary



Bayes' rule



Derived from the product rule:

- P(A, B) = P(A|B) * P(B) *# from definition of conditional probability*
- P(B, A) = P(B|A) * P(A) *# from definition of conditional probability*
- P(A, B) = P(B, A) *# since order is not important*

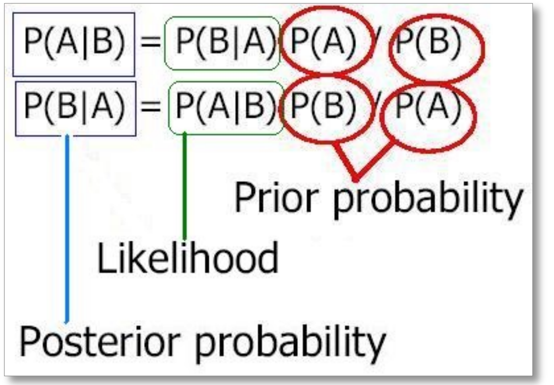
So…

$$\mathbf{P(A|B) = \frac{P(B|A) * P(A)}{P(B)}}$$

relates P(A|B) and P(B|A)

P(A,B) is probability of both A and B being true, so P(A,B) = P(B,A)

Useful for diagnosis!



- *C is a cause, E is an effect*:
  - P(C|E) = P(E|C) * P(C) / P(E)
- **Useful for diagnosis**:
  - E are (observed) effects and C are (hidden) causes,
  - Often have model for how causes lead to effects P(E|C)
  - We may have info (based on experience) on frequency of causes (P(C))
  - Which allows us to reason abductively from effects to causes (P(C|E))
  - Recall, abductive reasoning: from A => B and B, infer (maybe?) A

Example: meningitis and stiff neck

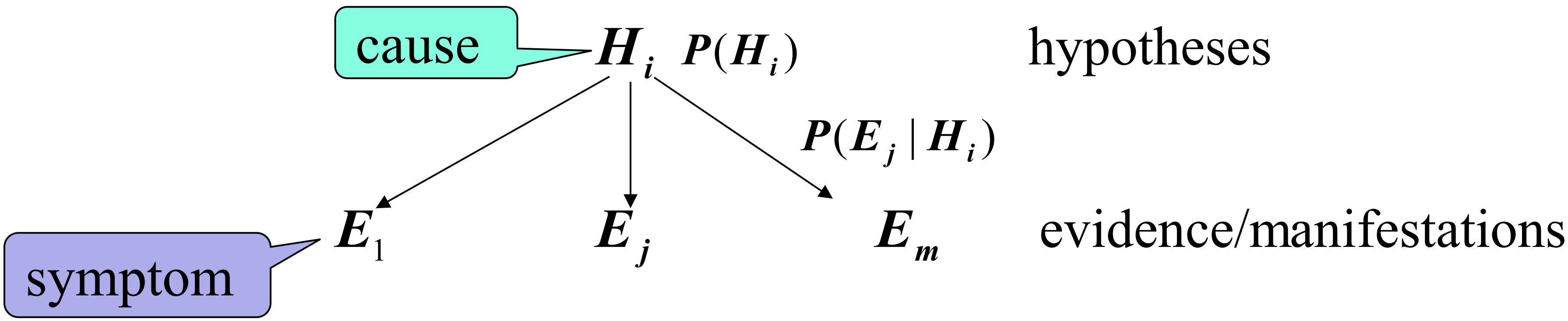
cause

symptom

- **Meningitis** (M) can cause **stiff neck** (S), though there are other causes too
- Use S as a *diagnostic symptom* & estimate **p(M|S)**
- Studies can estimate p(M), p(S) & p(S|M), e.g. p(S|M)=0.7, p(S)=0.01, p(M)=0.00002
- Harder to directly gather data on p(M|S)
- Applying Bayes' Rule:
  p(M|S) = p(S|M) * p(M) / p(S) = 0.0014

From multiple evidence to a cause

In the setting of diagnostic/evidential reasoning



– Know prior probability of hypothesis $P(H_i)$

conditional probability $P(E_j \mid H_i)$

– Want to compute the *posterior probability* $P(H_i \mid E_j)$

Bayes’ s theorem:

$$P(H_i \mid E_j) = P(H_i) * P(E_j \mid H_i) / P(E_j)$$

Bayesian diagnostic reasoning

- Knowledge base:
  - Evidence / manifestations: E_1, … E_m
  - Hypotheses / disorders: H_1, … H_n

    Note: E_j and H_i **binary**; hypotheses **mutually exclusive** (non-overlapping) & **exhaustive** (cover all possible cases)
  - Conditional probabilities: $P(E_j \mid H_i)$, $i = 1, \ldots n$; $j = 1, \ldots m$
- Cases (evidence for particular instance): $E_1, \ldots, E_l$
- Goal: Find hypothesis H_i with highest posterior
  - $\text{Max}_i\ P(H_i \mid E_1, \ldots, E_l)$

Bayesian diagnostic reasoning (2)

- Prior vs. posterior probability
  - Prior: probability before we know the evidence, e.g., 0.005 for having COVID)
  - Posterior: probability after knowing evidence, e.g., 0.9 if patient tests positive for COVID
- Goal: find hypothesis H_i with highest posterior
  - $\text{Max}_i \; P(H_i \mid E_1, \ldots, E_m)$
- Requires knowing joint evidence probabilities

$$P(H_i \mid E_1 \ldots E_m) = P(E_1 \ldots E_m \mid H_i)\, P(H_i) \,/\, P(E_1 \ldots E_m)$$

- Having many E_i is a big data collection problem!

Simplifying Bayesian diagnostic reasoning

- Having many E_i is a big data collection problem!
- Two ways to address this
- #1 use conditional independence to effect "causal reasoning" and eliminate some E_i
  - Knowing *LightLevel*, we can ignore *FullMoon* and *StreetLights* *when* predicting if alarm suggests Burglary
  - More on this later as Bayesian Believe Networks
- #2 Use a Naïve Bayes approximation that assumes evidence variables are all mutually independent

Simple Bayesian diagnostic reasoning

- Bayes' rule:

$$P(H_i \mid E_1 \dots E_m) = P(E_1 \dots E_m \mid H_i)\, P(H_i) \,/\, P(E_1 \dots E_m)$$

- Assume each evidence E_i is conditionally independent of the others, *given* a hypothesis H_i, then:

$$P(E_1 \dots E_m \mid H_i) = \prod_{j=1}^{m} P(E_j \mid H_i)$$

- If only care about relative probabilities for H_i, then:

$$P(H_i \mid E_1 \dots E_m) = \alpha\, P(H_i) \prod_{j=1}^{m} P(E_j \mid H_i)$$

Naive Bayes: Example

p(Wait | Cuisine, Patrons, Rainy?) =

$$= \alpha \bullet p(\text{Wait}) \bullet p(\text{Cuisine}|\text{Wait}) \bullet p(\text{Patrons}|\text{Wait}) \bullet p(\text{Rainy?}|\text{Wait})$$

$$= \frac{p(\text{Wait}) \bullet p(\text{Cuisine}|\text{Wait}) \bullet p(\text{Patrons}|\text{Wait}) \bullet p(\text{Rainy?}|\text{Wait})}{p(\text{Cuisine}) \bullet p(\text{Patrons}) \bullet p(\text{Rainy?})}$$

We can **estimate all of the parameters p(P) and p(C) just by counting from the training examples**

Naive Bayes: Analysis

- Naive Bayes is amazingly easy to implement (once you understand the math behind it)
- Naive Bayes can outperform many much more complex algorithms—it's a baseline that should be tried or used for comparison
- Naive Bayes can't capture interdependencies between variables (obviously)—for that, we need Bayes nets!

Bag of Words Classifier

classifier

$$\gamma\left(\begin{array}{|c|c|}\hline \text{seen} & 2 \\ \hline \text{sweet} & 1 \\ \hline \text{whimsical} & 1 \\ \hline \text{recommend} & 1 \\ \hline \text{happy} & 1 \\ \hline \ldots & \ldots \\ \hline \end{array}\right) = c$$

seen	2
sweet	1
whimsical	1
recommend	1
happy	1
...	...

classifier

Naïve Bayes (NB) Classifier

$$\text{argmax}_Y \, p(X \mid Y) * p(Y)$$

label text

Start with Bayes Rule

Naïve Bayes (NB) Classifier

label

each word

$$\text{argmax}_Y \prod_t p(X_t|Y) * p(Y)$$

Iterate through possible vocab words

Adopt naïve bag of words representation X_t

Assume position doesn't matter

Learning for a Naïve Bayes Classifier

Assuming V vocab types $w_1, \dots, w_V$ and L classes $c_1, \dots, c_L$(and appropriate corpora)

Learning for a Naïve Bayes Classifier

Assuming V vocab types $w_1, \ldots, w_V$ and L classes $c_1, \ldots, c_L$(and appropriate corpora)

Q: What parameters (values/weights) must be learned?

Learning for a Naïve Bayes Classifier

Assuming V vocab types $w_1, \dots, w_V$ and L classes $c_1, \dots, c_L$ (and appropriate corpora)

Q: What parameters (values/weights) must be learned?

A: $p(w_v | c_l), p(c_l)$

Learning for a Naïve Bayes Classifier

Assuming V vocab types $w_1, \ldots, w_V$ and L classes $c_1, \ldots, c_L$ (and appropriate corpora)

Q: What parameters (values/weights) must be learned?

A: $p(w_v | c_l), p(c_l)$

Q: How many parameters must be learned?

Learning for a Naïve Bayes Classifier

Assuming V vocab types $w_1, \ldots, w_V$ and L classes $c_1, \ldots, c_L$(and appropriate corpora)

Q: What parameters (values/weights) must be learned?

A: $p(w_v|c_l), p(c_l)$

Q: How many parameters must be learned?

A: $LV + L$

Learning for a Naïve Bayes Classifier

Assuming V vocab types $w_1, \dots, w_V$ and L classes $c_1, \dots, c_L$ (and appropriate corpora)

Q: What parameters (values/weights) must be learned?

A: $p(w_v|c_l), p(c_l)$

Q: How many parameters must be learned?

A: $LV + L$

Q: What distributions need to sum to 1?

Learning for a Naïve Bayes Classifier

Assuming V vocab types $w_1, \dots, w_V$ and L classes $c_1, \dots, c_L$(and appropriate corpora)

Q: What parameters (values/weights) must be learned?

A: $p(w_v | c_l), p(c_l)$

Q: How many parameters must be learned?

A: $LV + L$

Q: What distributions need to sum to 1?

A: Each $p(\cdot\ | c_l)$, and the prior

Multinomial Naïve Bayes: Learning

From training corpus, extract *Vocabulary*

Calculate $P(c_j)$ terms

For each c_j in C do

docs$_j$ = all docs with class $= c_j$

$$p(c_j) = \frac{|docs_j|}{\# \, docs}$$

Calculate $P(w_k \mid c_j)$ terms

Text$_j$ = single doc containing all *docs*$_j$

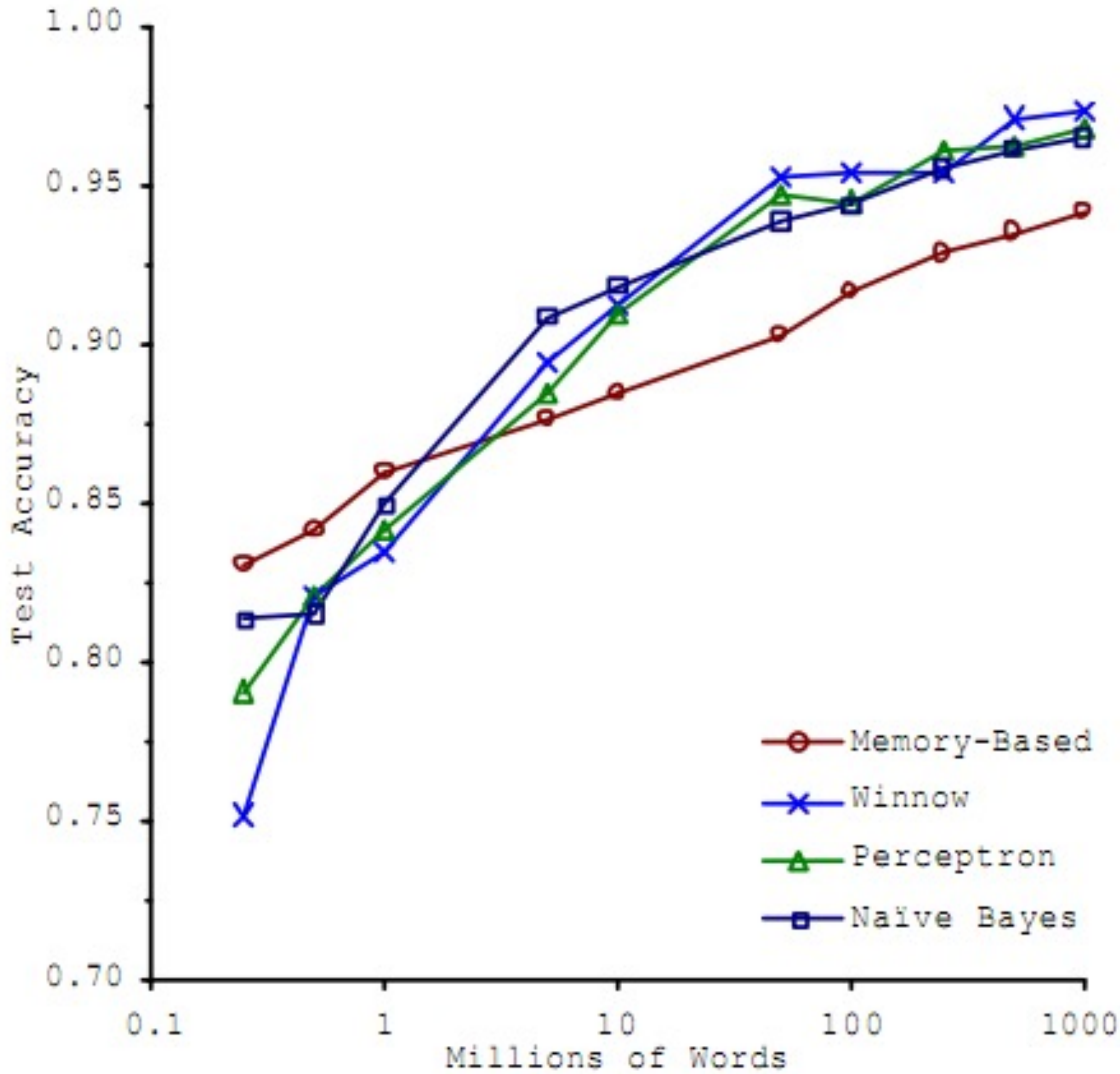
For each word w_k in *Vocabulary*

n_k = # of occurrences of w_k in *Text*$_j$

$$p(w_k \mid c_j) \propto \text{count(word } w_k \text{ in doc labeled with } c_j)$$

Naive Bayes: Analysis

- Naive Bayes is amazingly easy to implement (once you understand the math behind it)
- Naive Bayes can outperform many much more complex algorithms—it's a baseline that should be tried or used for comparison
- Naive Bayes can't capture interdependencies between variables (obviously)—for that, we need Bayes nets!



Brill and Banko (2001)

With enough data, the classifier may not matter

Naive Bayes: Analysis

- Naive Bayes is amazingly easy to implement (once you understand the math behind it)
- Naive Bayes can outperform many much more complex algorithms—it's a baseline that should be tried or used for comparison
- Naive Bayes can't capture interdependencies between variables (obviously)—for that, we need Bayes nets!

Limitations

- Can't easily handle **multi-fault situations** or cases where intermediate (hidden) causes exist:
  - Disease D causes syndrome S, which causes correlated manifestations M_1 and M_2
- Consider composite hypothesis $H_1 \wedge H_2$, where H_1 & H_2 independent. What's relative posterior?

$$P(H_1 \wedge H_2 \mid E_1, \ldots, E_l) = \alpha\, P(E_1, \ldots, E_l \mid H_1 \wedge H_2)\, P(H_1 \wedge H_2)$$

$$= \alpha\, P(E_1, \ldots, E_l \mid H_1 \wedge H_2)\, P(H_1)\, P(H_2)$$

$$= \alpha \prod^{l}_{j=1} P(E_j \mid H_1 \wedge H_2)\, P(H_1)\, P(H_2)$$

- How do we compute $P(E_j \mid H_1 \wedge H_2)$?

Limitations

- Assume H1 and H2 independent, given E1, …, El?
  - $P(H_1 \wedge H_2 \mid E_1, \ldots, E_l) = P(H_1 \mid E_1, \ldots, E_l)\, P(H_2 \mid E_1, \ldots, E_l)$
- Unreasonable assumption
  - Earthquake & Burglar independent, but *not* given Alarm:

    P(burglar | alarm, earthquake) << P(burglar | alarm)
- Doesn’t allow causal chaining:
  - A: 2017 weather; B: 2017 corn production; C: 2018 corn price
  - A influences C indirectly: A→ B → C
  - P(C | B, A) = P(C | B)
- Need richer representation for interacting hypoteses, conditional independence & causal chaining
- Next: Bayesian Belief networks!

Summary

- Probability a rigorous formalism for uncertain knowledge
- Joint probability distribution specifies probability of every atomic event
- Answer queries by summing over atomic events
- Must reduce joint size for non-trivial domains
- Bayes rule: compute from known conditional probabilities, usually in causal direction
- Independence & conditional independence provide tools
- Next: Bayesian belief networks