Bayes Classification

- Uncertainty & Probability
- Baye's rule
- Choosing Hypotheses- Maximum a posteriori
- Maximum Likelihood - Baye's concept learning
- Maximum Likelihood of real valued function
- Bayes optimal Classifier
- Joint distributions
- Naive Bayes Classifier

Uncertainty

- Our main tool is the probability theory, which assigns to each sentence numerical degree of belief between *0* and *1*
- It provides a way of summarizing the uncertainty

Variables

- Boolean random variables: cavity might be true or false
- Discrete random variables: weather might be sunny, rainy, cloudy, snow
  - *P(Weather=sunny)*
  - *P(Weather=rainy)*
  - *P(Weather=cloudy)*
  - *P(Weather=snow)*
- Continuous random variables: the temperature has continuous values

Where do probabilities come from?

- Frequents:
  - From experiments: form any finite sample, we can estimate the true fraction and also calculate how accurate our estimation is likely to be
- Subjective:
  - Agent’s believe
- Objectivist:
  - True nature of the universe, that the probability up heads with probability 0.5 is a probability of the coin

- Before the evidence is obtained; prior probability
  - *P(a)* the prior probability that the proposition is true
  - $P(cavity)=0.1$
- After the evidence is obtained; posterior probability
  - $P(a|b)$
  - The probability of a given that all we know is b
  - $P(cavity|toothache)=0.8$

Axioms of Probability

(Kolmogorov's axioms, first published in German 1933)

- All probabilities are between 0 and 1. For any proposition *a* $0 \le P(a) \le 1$
- $P(true)=1, \ P(false)=0$
- The probability of disjunction is given by

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

- Product rule

$$P(a \wedge b) = P(a \mid b)P(b)$$

$$P(a \wedge b) = P(b \mid a)P(a)$$

Theorem of total probability

If events $A_1, \ldots , A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$ then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

$$P(B) = \sum_{i=1}^{n} P(B, A_i)$$

Bayes's rule

- (Reverent Thomas Bayes 1702-1761)
  - He set down his findings on probability in "Essay Towards Solving a Problem in the Doctrine of Chances" (1763), published posthumously in the *Philosophical Transactions of the Royal Society of London*

$$P(b \mid a) = \frac{P(a \mid b)P(b)}{P(a)}$$

Diagnosis

- What is the probability of meningitis in the patient with stiff neck?

  - A doctor knows that the disease meningitis causes the patient to have a stiff neck in 50% of the time -> *P(s|m)*

  - Prior Probabilities:
    - That the patient has meningitis is 1/50.000 -> *P(m)*
    - That the patient has a stiff neck is 1/20 -> *P(s)*

$$P(m \mid s) = \frac{P(s \mid m) P(m)}{P(s)}$$

$$P(m \mid s) = \frac{0.5 * 0.00002}{0.05} = 0.0002$$

Normalization

$$P(y \mid x) = \frac{P(x \mid y)P(y)}{P(x)}$$

$$P(\neg y \mid x) = \frac{P(x \mid \neg y)P(\neg y)}{P(x)}$$

$$1 = P(y \mid x) + P(\neg y \mid x)$$

$$P(Y \mid X) = \alpha \times P(X \mid Y)P(Y)$$

$$\alpha \langle P(y \mid x), P(\neg y \mid x) \rangle$$

$$\alpha \langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle$$

Bayes Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- *P(h)* = prior probability of hypothesis h
- *P(D)* = prior probability of training data *D*
- *P(h|D)* = probability of *h* given *D*
- *P(D|h)* = probability of *D* given *h*

Choosing Hypotheses

- Generally want the most probable hypothesis given the training data
- **Maximum a posteriori** hypothesis h_{MAP}:

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$

$$= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D|h)P(h)$$

- If assume $P(h_i)=P(h_j)$ for all h_i and h_j, then can further simplify, and choose the
- **Maximum likelihood** (ML) hypothesis

$$h_{ML} = \arg\max_{h_i \in H} P(D|h_i)$$

Example

- Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result (+) in only 98% of the cases in which the disease is actually present, and a correct negative result (-) in only 97% of the cases in which the disease is not present

Furthermore, 0.008 of the entire population have this cancer

Suppose a positive result (+) is returned...

$P(cancer) = 0.008 \qquad P(\neg cancer) = 0.992$

$P(+|cancer) = 0.98 \qquad P(-|cancer) = 0.02$

$P(+|\neg cancer) = 0.03 \qquad P(-|\neg cancer) = 0.97$

$P(+|cancer) \cdot P(cancer) = 0.98 \cdot 0.008 = 0.0078$

$P(+|\neg cancer) \cdot P(\neg cancer) = 0.03 \cdot 0.992 = 0.0298$

$h_{MAP} = \neg cancer$

Normalization

$$\frac{0.0078}{0.0078+0.0298}=0.20745 \qquad \frac{0.0298}{0.0078+0.0298}= 0.79255$$

- The result of Bayesian inference depends strongly on the prior probabilities, which must be available in order to apply the method

Brute-Force Bayes Concept Learning

- For each hypothesis *h* in *H*, calculate the posterior probability

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} =_{h \in H} P(h|D)$$

- Given no prior knowledge that one hypothesis is more likely than another, what values should we specify for *P(h)*?
- What choice shall we make for *P(D|h)* ?

- Choose *P(h)* to be uniform distribution
- $$P(h) = \frac{1}{|H|}$$ for all *h* in *H*

- *P(D|h)=1* if *h* consistent with *D*
- *P(D|h)=0* otherwise

P(D)

$$P(D) = \sum_{h_i \in H} P(D \mid h_i) P(h_i)$$

$$P(D) = \sum_{h_i \in VS_{H,D}} 1 \cdot \frac{1}{|H|} + \sum_{h_i \notin VS_{H,D}} 0 \cdot \frac{1}{|H|}$$

$$P(D) = \frac{|VS_{H,D}|}{|H|}$$

- Version space $VS_{H,D}$ is the subset of consistent Hypotheses from H with the training examples in D

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

$$P(h|D) = \frac{0 \cdot P(h)}{P(D)} = 0$$ if h is inconsistent with D

$$P(h \mid D) = \frac{1 \cdot \frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}} = \frac{1}{|VS_{H,D}|}$$ if h is consistent with D

$$P(h|D) = \begin{cases} \frac{1}{|VS_{H,D}|} \text{if } h \text{ is consistent with } D \\ 0 \quad \text{otherwise} \end{cases}$$

Maximum Likelihood of real valued function

$$h_{ML} = arg \max_{h \in H} p(D|h)$$

$$= arg \max_{h \in H} \prod_{i=1}^{m} p(d_i|h)$$

$$= arg \max_{h \in H} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2}$$

- Maximize natural log of this instead...

$$h_{ML} = arg\max_{h \in H} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2$$

$$= arg\max_{h \in H} \sum_{i=1}^{m} -\frac{1}{2}\left(\frac{d_i - h(x_i)}{\sigma}\right)^2$$

$$= arg\max_{h \in H} \sum_{i=1}^{m} -\left(d_i - h(x_i)\right)^2$$

$$= arg\min_{h \in H} \sum_{i=1}^{m} \left(d_i - h(x_i)\right)^2$$

Bayes optimal Classifier

A weighted majority classifier

- What is he most probable classification of the new **instance** given the training data?
  - The most probable classification of the new instance is obtained by combining the prediction of *all hypothesis*, weighted by their *posterior probabilities*
- If the classification of new example can take any value v_j from some set V, then the probability $P(v_j|D)$ that the correct classification for the new **instance is v_j**, is just:

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Bayes optimal classification:

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

$$P(h_1|D) = .4, P(-|h_1) = 0, P(+|h_1) = 1$$

$$P(h_2|D) = .3, P(-|h_2) = 1, P(+|h_2) = 0$$

$$P(h_3|D) = .3, P(-|h_3) = 1, P(+|h_3) = 0$$

$$\sum_{h_i \in H} P(+|h_i)P(h_i|D) = .4 \qquad \sum_{h_i \in H} P(-|h_i)P(h_i|D) = .6$$

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = -$$

Gibbs Algorithm

- Bayes optimal classifier provides best result, but can be expensive if many hypotheses

- Gibbs algorithm:
  - Choose one hypothesis at random, according to $P(h|D)$
  - Use this to classify new instance

- Suppose correct, uniform prior distribution over *H*, then
  - Pick any hypothesis at random..
  - Its expected error no worse than twice Bayes optimal

Joint distribution

- A joint distribution for toothache, cavity, catch, *dentist‘s probe catches in my tooth :-(*

- We need to know the conditional probabilities of the conjunction of toothache and cavity
- What can a dentist conclude if the probe catches in the aching tooth?

$$P(cavity \mid toothache \wedge catch) = \frac{P(toothache \wedge catch \mid cavity) P(cavity)}{P(toothache \wedge cavity)}$$

- For *n* possible variables there are 2^n possible combinations

Conditional Independence

- Once we know that the patient has cavity we do not expect the probability of the probe catching to depend on the presence of toothache

$$P(catch \mid cavity \wedge toothache) = P(catch \mid cavity)$$
$$P(toothache \mid cavity \wedge catch) = P(toothache \mid cavity)$$

- Independence between a and b

$$P(a \mid b) = P(a)$$
$$P(b \mid a) = P(b)$$

$$P(a \wedge b) = P(a)P(b)$$

$$P(toothache, catch, cavity, Weather = cloudy) =$$

$$= P(Weather = cloudy)P(toothache, catch, cavity)$$

- The decomposition of large probabilistic domains into weakly connected subsets via conditional independence is one of the most important developments in the recent history of AI

- This can work well, even the assumption is not true!

A single cause directly influence a number of effects, all of which are conditionally independent

$$P(cause, effect_1, effect_2, \ldots effect_n) = P(cause)\prod_{i=1}^{n} P(effect_i \mid cause)$$

Naive Bayes Classifier

- Along with decision trees, neural networks, nearest nbr, one of the most practical learning methods

- When to use:
  - Moderate or large training set available
  - Attributes that describe instances are conditionally independent given classification
- Successful applications:
  - Diagnosis
  - Classifying text documents

Naive Bayes Classifier

- Assume target function *f: X* ➔ *V*, where each instance *x* described by attributes $a_1, a_2 .. a_n$
- Most probable value of *f(x)* is:

$$v_{MAP} = \operatorname*{arg\,max}_{v_j \in V} P(v_j | a_1, a_2 \ldots a_n)$$

$$v_{MAP} = \operatorname*{arg\,max}_{v_j \in V} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

$$= \operatorname*{arg\,max}_{v_j \in V} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

v_{NB}

- Naive Bayes assumption:

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

- which gives

Naive Bayes classifier: $v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$

Naive Bayes Algorithm

- For each target value v_j
- $\hat{P}(v_j)$ ← estimate $P(v_j)$
- For each attribute value a_i of each attribute a
- $\hat{P}(a_i|v_j)$ ← estimate $P(a_i|v_j)$

$$v_{NB} = \arg\max_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$$

Training dataset

Class:
C1:buys_computer='yes'
C2:buys_computer='no'

Data sample:

X =
(age<=30,
Income=medium,
Student=yes
Credit_rating=Fair)

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
30…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no

Naïve Bayesian Classifier: Example

- Compute P(X|C_i) for each class

P(buys_computer=„yes“)=9/14

P(buys_computer=„no“)=5/14

P(age=“<30” | buys_computer=“yes”) = 2/9=0.222
P(age=“<30” | buys_computer=“no”) = 3/5 =0.6
P(income=“medium” | buys_computer=“yes”)= 4/9 =0.444
P(income=“medium” | buys_computer=“no”) = 2/5 = 0.4
P(student=“yes” | buys_computer=“yes)= 6/9 =0.667
P(student=“yes” | buys_computer=“no”)= 1/5=0.2
P(credit_rating=“fair” | buys_computer=“yes”)=6/9=0.667
P(credit_rating=“fair” | buys_computer=“no”)=2/5=0.4

- X=(age<=30 ,income =medium, student=yes,credit_rating=fair)

P(X|C_i) : P(X|buys_computer=“yes”)= 0.222 x 0.444 x 0.667 x 0.0.667 =0.044
P(X|buys_computer=“no”)= 0.6 x 0.4 x 0.2 x 0.4 =0.019

P(X|C_i)*P(C_i) : P(X|buys_computer=“yes”) * P(buys_computer=“yes”)=0.028
P(X|buys_computer=“no”) * P(buys_computer=“no”)=0.007

- X belongs to class “buys_computer=yes”

- Conditional independence assumption is often violated
- ...but it works surprisingly well anyway

Estimating Probabilities

- We have estimated probabilities by the fraction of times the event is observed to n_c occur over the total number of opportunities n
- It provides poor estimates when n_c is very small
- If none of the training instances with target value v_j have attribute value a_i?
  - n_c is 0

- When n_c is very small:

$$\hat{P}(a_i|v_j) = \frac{n_c + mp}{n+m}$$

- n is number of training examples for which $v=v_j$
- n_c number of examples for which $v=v_j$ and $a=a_i$
- p is **prior** estimate
- m is weight given to prior (i.e. number of ``virtual'' examples)

$$v_{NB} =_{v_j \in V} P(v_j) \prod_i \hat{P}(a_i|v_j)$$

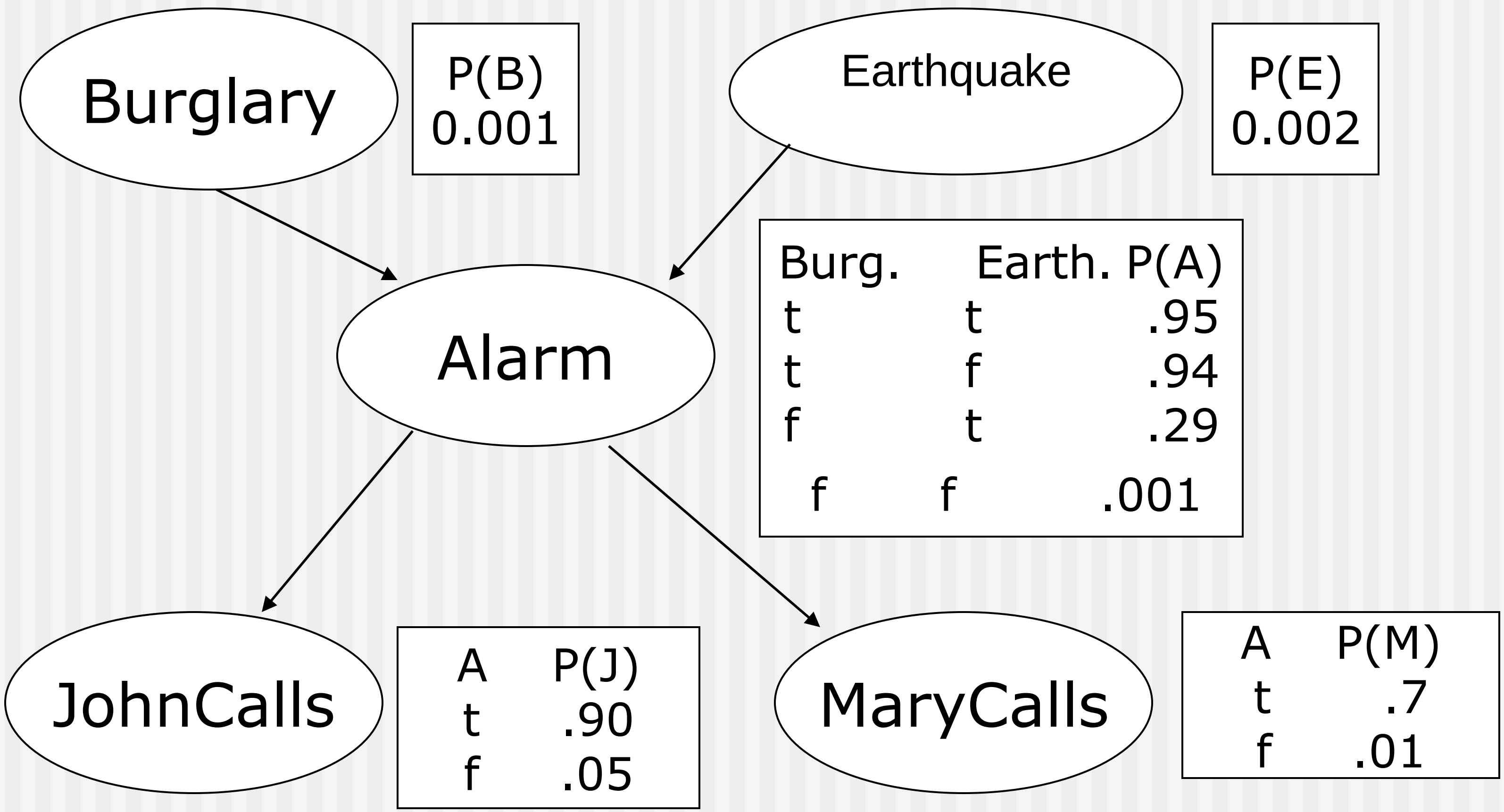
Naïve Bayesian Classifier: Comments

- Advantages :
  - Easy to implement
  - Good results obtained in most of the cases
- Disadvantages
  - Assumption: class conditional independence , therefore loss of accuracy
  - Practically, dependencies exist among variables
  - E.g., hospitals: patients: Profile: age, family history etc

    Symptoms: fever, cough etc., Disease: lung cancer, diabetes etc
  - Dependencies among these cannot be modeled by Naïve Bayesian Classifier
- How to deal with these dependencies?
  - Bayesian Belief Networks

- Uncertainty & Probability
- Baye's rule
- Choosing Hypotheses- Maximum a posteriori
- Maximum Likelihood - Baye's concept learning
- Maximum Likelihood of real valued function
- Bayes optimal Classifier
- Joint distributions
- Naive Bayes Classifier

Bayesian Belief Networks



Thank you !!!!

Any Questions ????

Utkarsh Kulshrestha

(kuls.utkarsh1205@gmail.com)