5-8-2012

Bayesian and Related Methods: Techniques Based on Bayes' Theorem

Mehmet Vurkaç
Portland State University

Bayesian and Related Methods

Techniques based on Bayes' Theorem

Mehmet Vurkaç, 5/18/2012

Outline

- Introduction & Definitions
- Bayes’ Theorem
- MAP Hypothesis & Maximum Likelihood
- Bayes Optimal & Naïve Bayes Classifiers
- Bayesian Decision Theory
- Bayesian Belief Nets
- Other “Famous” Applications

Introduction

- Motivation for Talk
- Numerical way to weigh evidence
- Medicine, Law, Learning, Model Evaluation
- Outperform other methods?
- Priors (Base Rates)
- Computationally expensive

Machine Learning

- Space of hypotheses
- Find “best”
  - Most likely true / underlying
  - Given data or domain knowledge

Definitions

- $P(h) \triangleq$ initial prob. that *h* holds
- $P(D) \triangleq$ likelihood of observing a set of data, *D*
- $P(D|h) \triangleq$ likelihood of observing *D* given some set of circumstances (universe/context) where *h* holds

ML goal is to rate and select hypotheses:

- $P(h|D) \triangleq$ probability that *h* holds GIVEN that *D* were observed

Conditional Prob. & Bayes' Theorem

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

Rearranging:

- $P(B|A) = \dfrac{P(A|B)P(B)}{P(A)}$
- $posterior = \dfrac{likelihood \times prior}{evidence}$

Bayes' Theorem

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

Maximum-*a posteriori* Hypothesis

$$h_{MAP} = \underset{h \in H}{argmax} \frac{P(D|h)P(h)}{P(D)}$$

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

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

Maximum-Likelihood Hypothesis

$$h_{ML} = \underset{h \in H}{argmax}\, P(D|h)$$

Example: Cancer test

- Existing data
- Imperfect test
- New patient gets a positive result.
- Should we conclude s/he has this cancer?

Example: Cancer test

- Test gives true positives in 98% of cases of cancer.
- Test gives true negatives in 97% in cases without cancer.
- 0.8% of population on record has this cancer.

Example: Inventory of Information

- $P(cancer) = 0.008$
- $P(\neg cancer) = 0.992$
- $P(+|cancer) = 0.980$
- $P(-|cancer) = 0.020$
- $P(+|\neg cancer) = 0.030$
- $P(-|\neg cancer) = 0.970$

Goal: Find MAP hypothesis

- "$P(cancer|+)$" = $P(+|cancer)P(cancer)$ = $(0.980)(0.008) = 0.0078$
- "$P(\neg cancer|+)$" = $P(+|\neg cancer)P(\neg cancer)$ = $(0.0030)(0.992) = 0.0298$
- $0.0298 > 0.0078$; diagnosis: no cancer
- And how likely is that to be true?

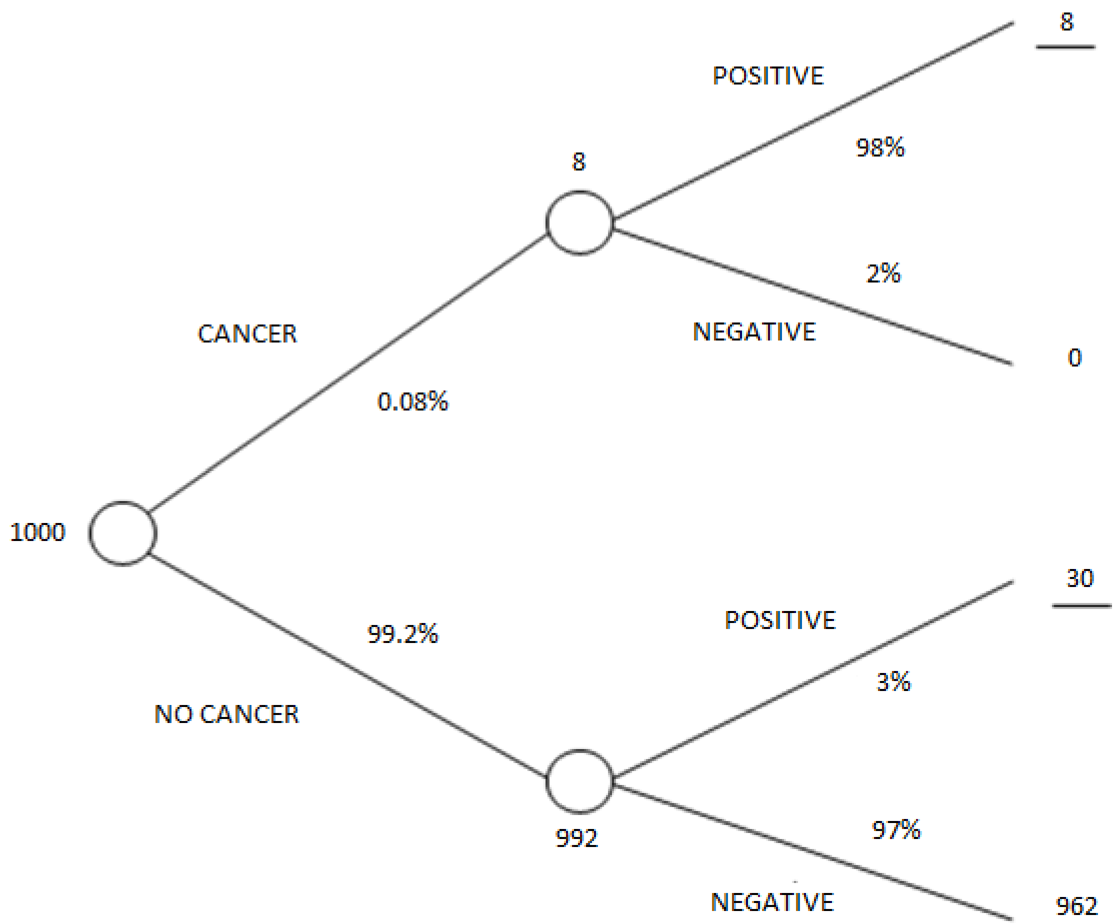
Human Aspect

$$P(cancer|+) = \frac{P(+|cancer)P(cancer)}{P(+)}$$

$$= \frac{P(+|cancer)P(cancer)}{P(+|cancer)P(cancer) + P(+|\neg cancer)P(\neg cancer)}$$

$$= 0.21$$

Example: Probability Tree



Bayes Optimal Classifier

- Adds the ensemble of hypotheses to MAP.
  - Contexts
- Assume we know:
  - $P(h_1|D) = 0.40$
  - $P(h_2|D) = 0.30$
  - $P(h_3|D) = 0.30$
- h_1 is the MAP hypothesis, so conclude +?
- $P(+) = 0.40 \qquad P(–) = 0.60$

Bayes Optimal Classifier

- Classifying data into one of many categories
- Under several hypotheses
- Categories: $v_1, v_2, v_3, \ldots, v_i, \ldots, v_m$
- Hypotheses: $h_1, h_2, h_3, \ldots, h_j, \ldots, h_n$

$$P(v_i|D) = \sum_{h_j \in H} P(v_i|h_j) P(h_j|D)$$

- and

$$\underset{v_i \in V}{argmax} \sum_{h_j \in H} P(v_i|h_j) P(h_j|D)$$

Bayes Optimal (BOC) & Gibbs

- No other method can outperform BOC *on* average.
- BOC must calculate every posterior, and compare them all.
- Gibbs
  - picks one *h* from *H* for each instance
  - weighted similarly to roulette wheel in GAs

Working with Features

- Typically, we work with multiple features
- Mathematically the same as multiple hypotheses.
- Vector of features: $x_1^p, x_2^p, x_3^p, \ldots, x_j^p, \ldots, x_n^p$
- Categories: c_i
- To make a MAP decision given a feature vector

$$c_{MAP} = \underset{c_i \in C}{argmax} \, P(c_i|\vec{x}^p)$$

Features & MAP

- which, by Bayes' Theorem, equals

$$\underset{c_i \in C}{argmax} \frac{P(x_1^p = a_1, \dots, x_j^p = a_j, \dots, x_n^p = a_n | c_i) P(c_i)}{P(x_1^p = a_1, \dots, x_j^p = a_j, \dots, x_n^p = a_n)}$$

- We can use the MAP simplification to get

$$c_{MAP} = \underset{c_i \in C}{argmax} P(\vec{x}^p = \vec{a}_j | c_i) P(c_i)$$

MAP Computational Cost

- To estimate these probabilities, we need numerous copies of every feature-value combination for each category.
  - many examples

    ×
  - feature combinations

    ×
  - categories

Reducing Computational Cost, Naively

- Assume features are independent.
    - P(observing a vector)

      becomes
    - product of P(observing each feature)

$$c_{NB} = \underset{c_i \in C}{\operatorname{argmax}} P(c_i) \prod_j P\left(x_j^p = a_j \middle| c_i\right)$$

- Rarely true!

Reducing Computational Cost, Naively

- Assume features are independent.
  - P(observing a vector)

    becomes
  - product of P(observing each feature)

$$c_{NB} = \underset{c_i \in C}{argmax} P(c_i) \prod_j P(x_j^p = a_j | c_i)$$

- Rarely true!

Quick Naïve-Bayes Example

- Student deciding what to do
    - Invited to a party: Y / N
    - Deadlines: Urgent / Near / None
    - Lazy: Y / N
    - Output classes: PARTY, HW, TV, BARS

Example: The Data

Deadlines?	Invited?	Lazy?	DECISION
Urgent	Y	Y	PARTY
Urgent	N	Y	HW
Near	Y	Y	PARTY
None	Y	N	PARTY
None	N	Y	BARS
None	Y	N	PARTY
Near	N	N	HW
Near	N	Y	TV
Near	Y	Y	PARTY
Urgent	N	N	HW
Near	N	N	BARS
None	Y	Y	TV
None	N	N	BARS
Urgent	N	N	HW
Near	Y	N	PARTY
None	N	N	BARS
Urgent	Y	Y	HW
None	Y	Y	TV
None	N	Y	TV
Urgent	Y	N	PARTY

Example: The Data

- “Probabilities”
  - $P(HW) = 5/20$
  - $P(PARTY) = 7/20$
  - $P(Invited) = 10/20$
  - $P(Lazy) = 10/20$
  - $P(PARTY|Lazy) = 3/10$
  - $P(Lazy|PARTY) = 3/7$

Classify a new instance

- Urgent / Invited / Lazy
  - P(decidePARTY) =

    $P(PARTY) \times P(Urgent|PARTY) \times P(Invited|PARTY) \times P(Lazy|PARTY)$

    $= (7/20) \times (2/7) \times (7/7) \times (3/7) = 0.042857\ldots$
  - $P(decideHW) = (5/20) \times (4/5) \times (1/5) \times (2/5) = 0.016$
  - $P(decideBARS) = (4/20) \times (0/4) \times (0/4) \times (1/4) = 0$
  - $P(decideTV) = (1/10) \times (0/1) \times (0/1) \times (1/1) = 0$

Bayesian Decision Theory

- Errors don't carry the same risk.
  - Loss penalties for decisions with risk
- We can also have an action of *not deciding*.
- Categories: c_i
- Actions: $\alpha_1, \alpha_2, \ldots, \alpha_k, \ldots, \alpha_a$
- Loss function: $\lambda_{ki} = \lambda(\alpha_k | c_i)$
- Conditional risk is expected loss for an action:

$$R(\alpha_k|\vec{x}) = \sum_{i=1}^{c} \lambda(\alpha_k|c_i) P(c_i|\vec{x})$$

- This time, argmin over the actions…

Minimax, Neyman-Pearson, ROC

A risky decision may need be taken under different conditions, different priors:

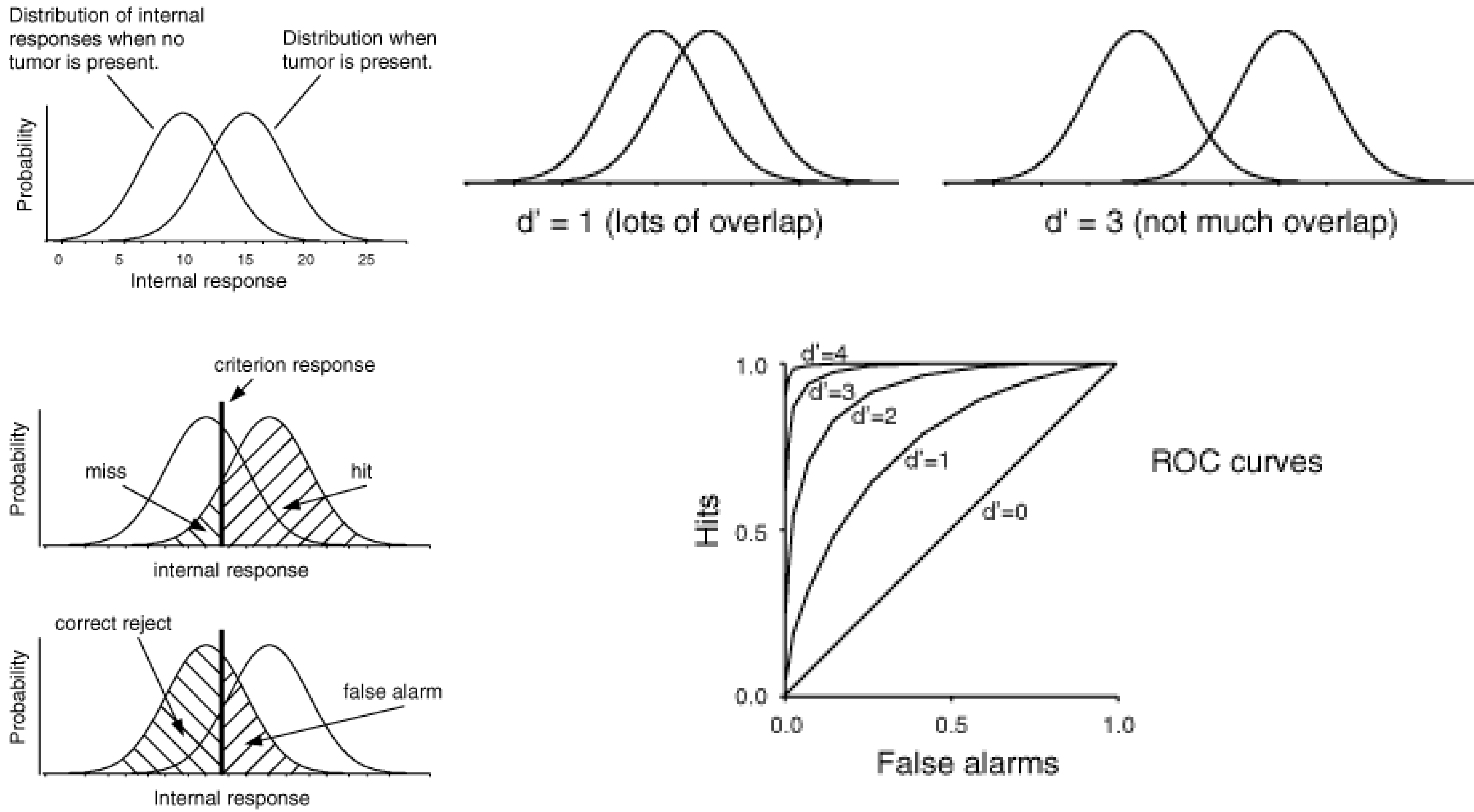
- Factories in different locations
- Seasons for biological studies
- Strategies for different competitor actions

- Design a classifier to minimize worst-case risk.
- Minimize overall risk subject to a constraint.
- In detecting a small stimulus, judge the quality of a threshold choice.

Receiver Operating Characteristic

- Plot hits (true positives) against false alarms.
- For choices of threshold, the same data give different curves.
- The areas under ROC curves correspond to a ranking of the probabilities that each threshold will allow correct identification of the small stimulus.

Receiver Operating Characteristic



http://www-psych.stanford.edu/~lera/psych115s/notes/signal/

Bayesian Belief Nets

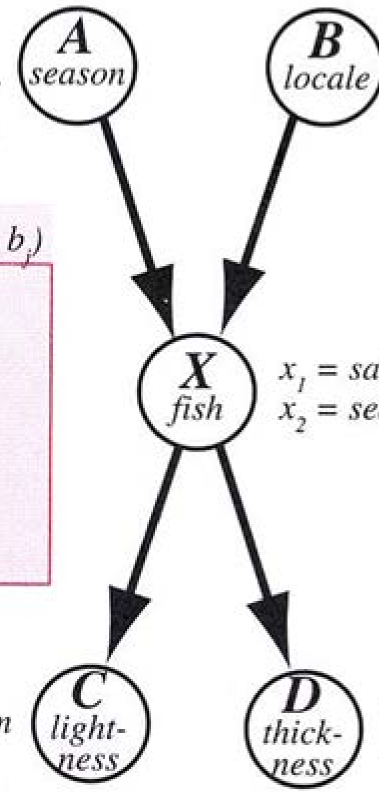
- Probabilistic reasoning
  - Using directed acyclic graphs
- Variables determine state of a system.
  - Some are causally related; some are not.
- Specified in conditional-probability tables
  - associated with each node (variable)
- Classification of caught fish (Duda, Hart, and Stork)

Bayesian Belief Nets

$P(a)$

$P(a_1)$	$P(a_2)$	$P(a_3)$	$P(a_4)$
0.25	0.25	0.25	0.25

a_1 = winter
a_2 = spring
a_3 = summer
a_4 = autumn



b_1 = north Atlantic
b_2 = south Atlantic

$P(b)$

$P(b_1)$	$P(b_2)$
0.6	0.4

$P(x|a,b)$

	$P(x_1 \vert a_i, b_j)$	$P(x_2 \vert a_i, b_j)$
a_1, b_1	0.5	0.5
a_1, b_2	0.7	0.3
a_2, b_1	0.6	0.4
a_2, b_2	0.8	0.2
a_3, b_1	0.4	0.6
a_3, b_2	0.1	0.9
a_4, b_1	0.2	0.8
a_4, b_2	0.3	0.7

x_1 = salmon
x_2 = sea bass

$P(c|x)$

	$P(c_1 \vert x_k)$	$P(c_2 \vert x_k)$	$P(c_3 \vert x_k)$
x_1	0.6	0.2	0.1
x_2	0.2	0.3	0.5

c_1 = light
c_2 = medium
c_3 = dark

d_1 = wide
d_2 = thin

$P(d|x)$

	$P(d_1 \vert x_k)$	$P(d_2 \vert x_k)$
x_1	0.3	0.7
x_2	0.6	0.4

Duda, Hart, Stork: Pattern Classification

Other Applications

- Bayesian learning is recursive
  - Spam filters that continue to learn after being deployed
  - Scientific investigation: new data update models
- HMM: Time-dependent BBN with unknown Markov state
- Viterbi: Most likely sequence of states
- Kalman: Next-state prediction, observation, correction by weighting the error computation with current trust in predictions – updated after more observations.
- PNN: kernel neural net implements MAP.
- The list goes on.

Bayes' Theorem

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

$$posterior = \frac{likelihood \ \times prior}{evidence}$$

References

- Mitchell, M., (class notes), PSU CS 410/510 TOP: Machine Learning, Portland State University, Winter Term, 2005.
- Profs. Don Moor, David Weber, and Peter Nicholls, *Knowledge, Rationality and Understanding*, University Studies, Portland State University, 2004–2007.
- Marsland, S., *Machine Learning: An Algorithmic Perspective*, Chapman & Hall/CRC Press/Taylor & Francis Group, Boca Raton, FL, 2009.
- Duda, R. O., Hart, P. E., and Stork, D. G., *Pattern Classification*, Wiley-Interscience/John Wiley & Sons, Inc., New York, NY, 2001.

References (cont.)

- Temperley, D., *Music and Probability*, The MIT Press, Cambridge, MA, 2007.
- Hawkins, J., and Blakeslee, S., *On Intelligence*, Macmillan, 2005.
- Kapur, J. N. and Kesavan, H. K., *Entropy Optimization Principles with Applications*, Academic Press, San Diego, 1992.
- Hopley, L., and van Schalkwyk, J., The Magnificent ROC, http://www.anaesthetist.com/mnm/stats/roc/Findex.htm
- Heeger, D., and Boroditsky, L., Signal Detection Theory Handout, http://www-psych.stanford.edu/~lera/psych115s/notes/signal/, 1998.

Discussion