Decisions and Computation

Dr. Wray Buntine
Heuristicrats Research, Inc.

wray@Heuristicrat.COM
http://www.Heuristicrat.COM/wray/

1678 Shattuck Avenue, Suite 310 • Berkeley, CA, 94709-1631
Tel: +1 (510) 845-5810 • Fax: +1 (510) 845-4405

Section in the tutorial at *Maximum Entropy and Bayesian Methods*, Sante Fe, New Mexico, June 31st, 1995.

Basic operations

marginalization : $p(X) = \sum_Y p(X,Y)$

- used from right to left to eliminate "nuisance" variables not required later
- used from left to right to introduce "hidden" variables

conditioning : $p(X \mid Y) = p(X,Y) / p(Y)$

- used to incorporate evidence (in this case Y)
- used for hypothetical reasoning, "what if we knew Y?"

factorization : $p(X,Y) = p(X)\ p(Y|X)$

independence : $p(X|Y,Z) = p(X|Z)$ assuming Y indep. of X given Z

- used in problem decomposition and simplification

expectation : Expected-Value $U(X) = \sum_X U(X)\ p(X)$

- used to calculate expected utilities, averages, etc.

maximization : Max_d Expected-Value U(X,d)

- used to maximize utility, choose best decision

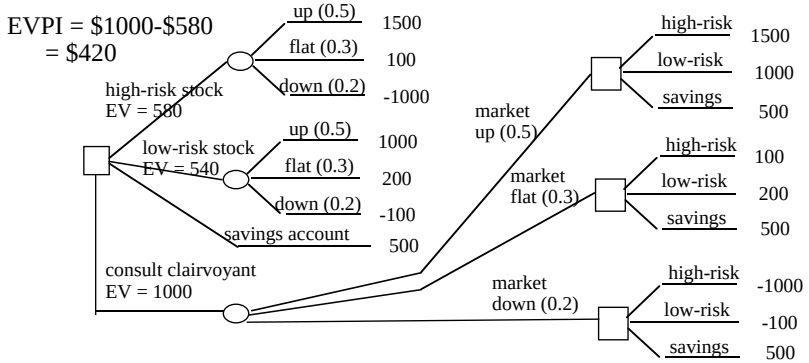
NB. and they all apply recursively; recursive maximization leads to dynamic programming; splicing recursive maximization and conditioning is the hardest, e.g. two-arm bandit, learning to balance a pole

Value of information (see Clemen, '90)

basic framework for managing the collection of information, sense acts, etc.

We are about to make an investment decision. What would be the value to us of knowing the truth about future stock market activity?

Expected value of perfect information about X: What would be our increase in utility from having a clairvoyant inform us about the true value for X? (Current utility is based on our belief of what X might be.)



The value of computation

basic framework for managing uncertainty/complexity in computation

e.g. Monte Carlo sampling with N cases $\{ x_1, x_2, ..., x_N \}$, should you sample a new case, x_{N+1}, or stop and make do with the sample you have?

cost of sampling a single case = T seconds

current variance of sample average

$$s^2 = \left(\sum_i x_i^2 - N \left(\sum_i x_i\right)^2 \right) / (N-1)$$

approximate decrease in variance = $s^2 \left(1 - (N-1)/N \right) = s^2 / N$

i.e. diminishing returns as $N \to \infty$

value of computation whats the value (in terms of change in utility) of doing the extra computation?

rate cost of a single second = S *how do we assess these?*

rate cost of increasing mean square error of estimate = M

value of computation $\approx M s^2 / N - ST$

these computations can be embedded in scheduling and optimization systems, etc.

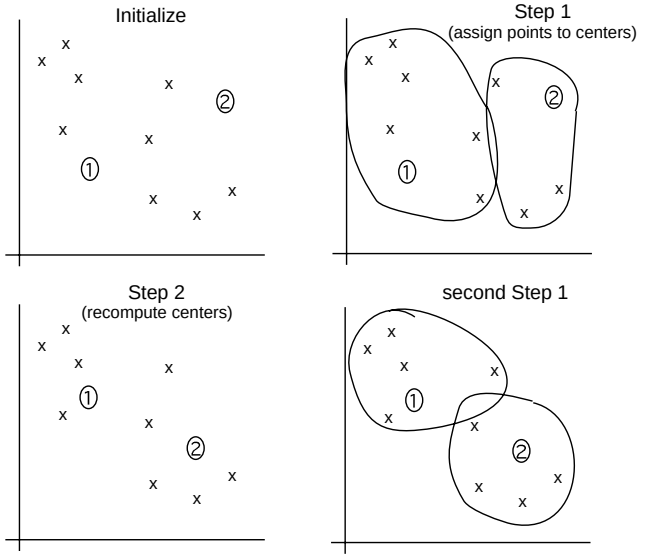
Mixture models

- Introduce an extra variable c (the mixing variable):

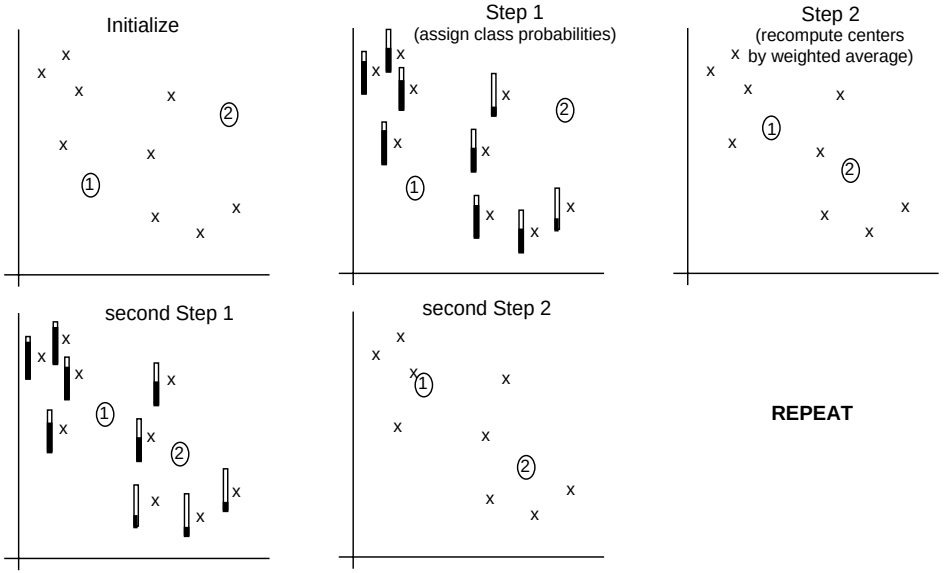
  $p(X) = \sum_c p(c)\, p(X|c)$

  where p(c) and p(X|c) are suitably tractible.
- Are ubiquitous in data analysis:
  - missing values in other problems, e.g. in curve-fitting data
  - latent or hidden variables, e.g., medical {\em syndromes}.
  - unsupervised learning and clustering,& hidden Markov models
  - Supervised learning and multivariate splits in trees, robust regression
  - non-parametric density estimation (i.e., equivalent to Kernel density estimation and nearest neighbor).
  - rule-based systems with multi-firing probabilistic rules.
- Examples over page are for: c is binary, p(c) is Bernoulli, p(X|c) is a 2-D Gaussian with unit covariance matrix.

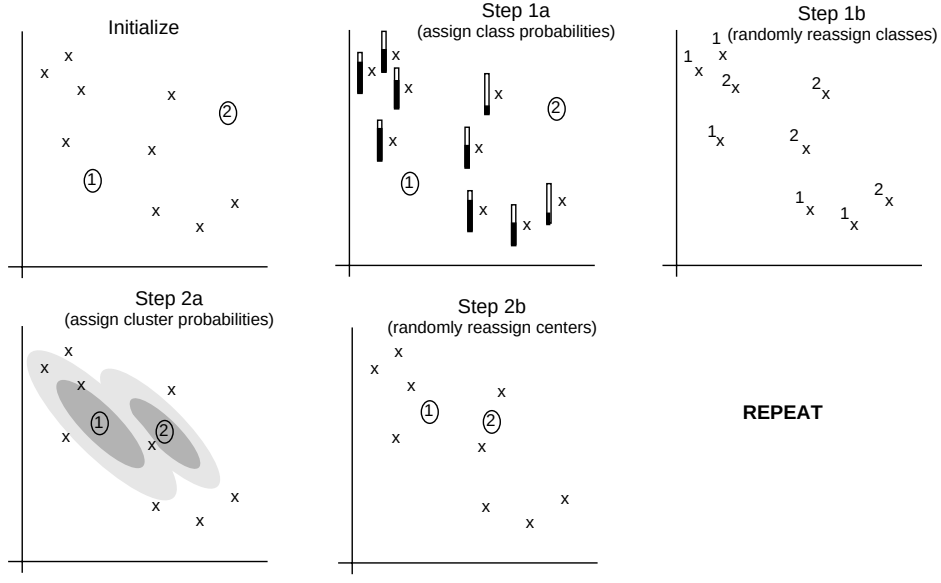
K-means algorithm in 2-D clustering



EM algorithm in 2-D clustering



Gibbs sampling in 2-D clustering



Algorithms

☞ Many probabilistic algorithms are closely related to one of the following:

Maximum A Posteriori (MAPs): via local search, gradient descent, maximum likelihood, MDL, minimum K-L

Exponential family or 2nd order exponential: linear regression, estimating Gaussian, Poisson or multinomial parameters (see any text)

NB. many other algrithms use these routines in their inner core

Monte Carlo Markov Chains (MCMCs): simulation to estimate expected values; Gibbs sampling and the Metropolis algorithm are variations; used in more challenging conputation where other algorithms wont apply (see any advanced text)

Algorithms, cont.

Variable clustering: recursive marginalization and conditioning on discrete/ Gaussian Bayesian networks; used in first order inference for diagnosis, and to simplify calculations for Bayesian classifiers (Shachter, Andersen, Szolovits, 94)

Exact algorithms for iterative updating: Kalman filters, forward-backward algorithm for HMMs, Viterbi algorithm, each are instances of Bayesian network clustering algorithms

Expectation-Maximization (EM): for estimation, unsupervised learning or clustering; Baum-Welch is a variation of this using the above exact routines; EM is a deterministic version of Gibbs sampling

Laplace's method: approximate Bayes factors, marginals and expected values (Tanner,1993; Kass and Raftery, 1995)

Large sample, parallel, and incremental versions: most algorithms are easily adapted for large sample problems, for incremental updating, and for parallel algorithms, e.g., sub-sampling, data parallelism, independent parallel search/ sampling, restarting search at last optima

Multinomial Priors

$\theta <$					$1-\theta <$				**grouped cells**
θ_1	θ_2	θ_3	•••	θ_{5000}	θ_{5001}	•••	θ_{9999}	θ_{10000}	**cells**

- mutually exclusive and exhaustive set of cells, so $\sum_{i \le 10000} \theta = 1$
- each cell has a probability of occurrence; i.e. a multinomial distribution

e.g. leaves in a decision tree, phonemes in a speech system, words in a language model, faults in a diagnosis system, cells in a Bayes network probability table, ...

Whats a "good" prior for the multinomial ?

Multinomial Priors, cont.

$\theta <$					$1-\theta <$				**grouped cells**
θ_1	θ_2	θ_3	•••	θ_{5000}	θ_{5001}	•••	θ_{9999}	θ_{10000}	**cells**

The textbook "non-informative" prior is Jeffreys' prior $\propto \prod_{i \le 10000} \theta_i^{-0.5}$

(Although some text books suggest as many as 4 different alternatives.)

How do we interpret/understand this ?

- look at expected values:

average $\theta_i = 1/10000$; std. dev. $\theta_i \approx 1.4/10000$ **!!**

average $\theta < = 0.5$; std. dev. $\theta < \approx 0.007$

☞ **The "non-informative" prior is highly informative about the grouped cells!!**

Aside: Non-informative Priors

Jeffreys' non-informative prior: $\pi(\theta) \propto \iota(\theta)^{1/2}$

where $\iota(\theta)$ is the determinant of the Fisher Information matrix.

NB 1. Jeffreys prior is known to be poor in certain contexts, such as:

- Gaussian with unknown mean and variance
- 10,000-dimensional multinomial $(\theta_1, \theta_2, \theta_3, \ldots \theta_{10,000})$; the marginal prior on $\sum_{i<5,000} \theta_i$ approaches a delta function so is highly informative

NB 2. Jeffreys argument is based on invariance; its poor because using measure theory there are infinitely many other priors with similar properties

See the discussion on "reference" priors in Bernardo and Smith, 1994. In general, "reference" priors for multi-dimensional spaces are difficult.