Bayesian Classification: AutoClass

Fanchao Meng
fcmeng@cis.udel.edu
Li Ren
renli@udel.edu

Computer and Information Sciences Department
University of Delaware

March 6, 2014

Outline

- Objective
- What do we use?
- What is Bayesian Classification?
- What is Finite Mixture Model?
- What to learn?
- How to learn?
- Attribute models & problems
- Case study
- What is more…

Objective

Find classes given data!

- Number of classes
- Relationships between classes
- Description of each class

What do we use?

Bayesian Classification: AutoClass

- Bayesian version Finite Mixture Model
- Parameters Search & Estimation

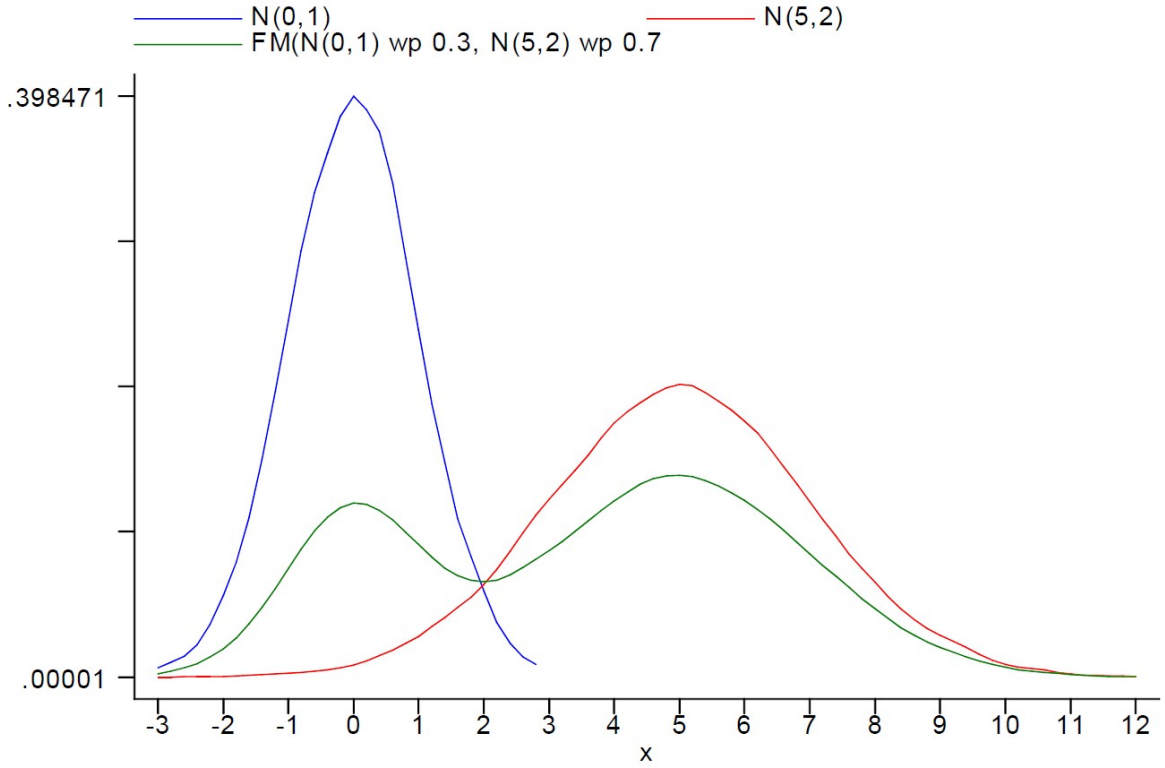
What is Bayesian Classification?

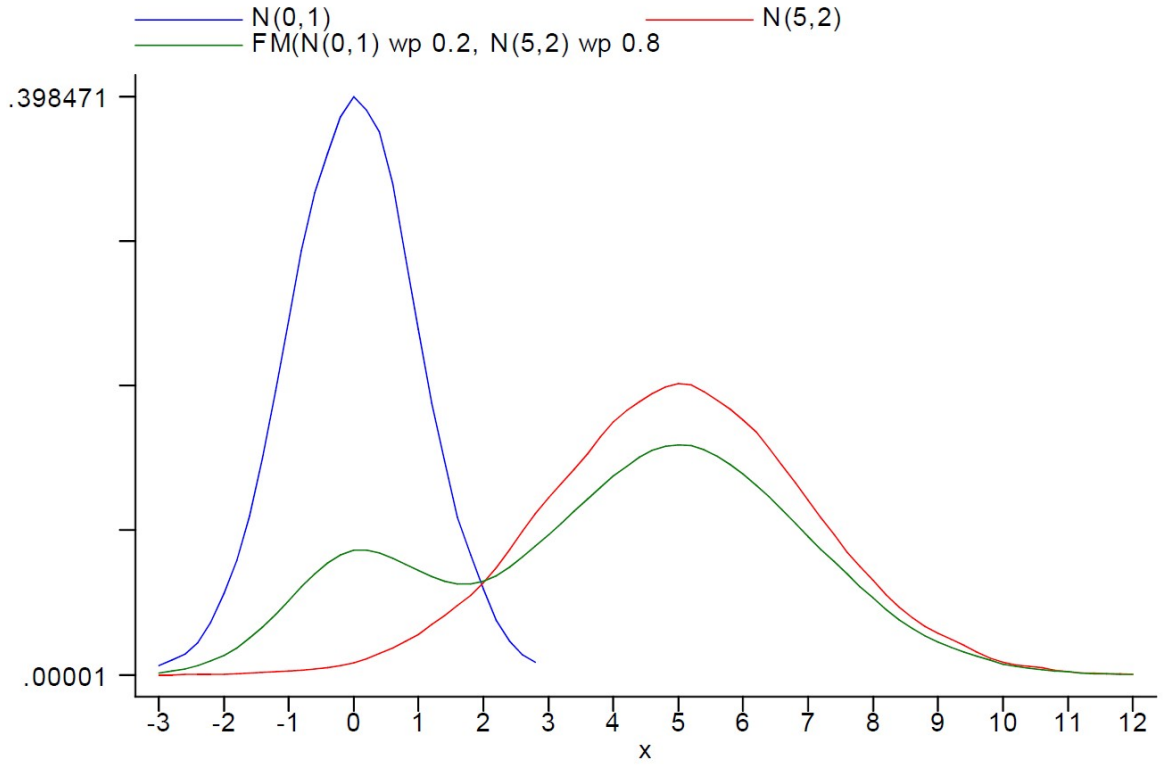
$$p(H|D) = \frac{p(H)p(D|H)}{p(D)}$$

H	hypothesis (number & descriptions of classes)
D	data
$p(H)$	prior probability of H
$p(D)$	prior probability of D
$p(D\|H)$	likelihood of D
$\boldsymbol{p(H\|D)}$	**posterior probability of H given D**

[12]

What is Finite Mixture Model?





[1]

What is Finite Mixture Model? (cont.)

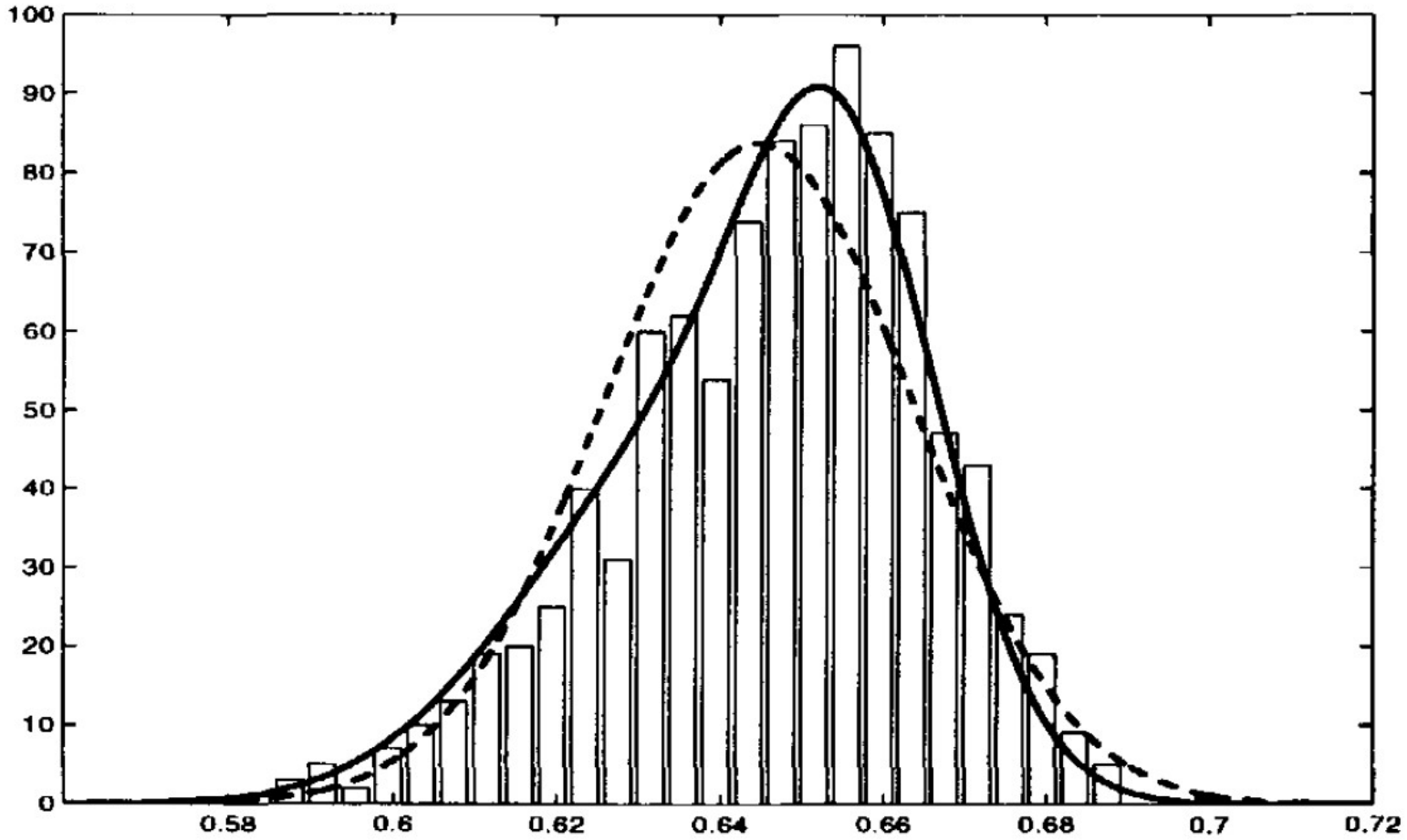


Fig. 1.1 Plot of forehead to body length data on 1000 crabs and of the fitted one-component (dashed line) and two-component (solid line) normal mixture models.

[2][4]

What is Finite Mixture Model? (cont.)

$\boldsymbol{X} = X_1, \ldots, X_I$ — I data instances X_i

$\overrightarrow{X_i} = \{X_{i1}, \ldots, X_{iK}\}$ — ith instance with K attributes

$\boldsymbol{T}_c, \boldsymbol{T} = T_1, \ldots, T_J$ — formal p.d.f. (without concrete parameters)
$\boldsymbol{T}_c$ is for mixture probabilities
T_i is for each component (class) [11]

$\boldsymbol{V} = \vec{V}_1, \ldots, \vec{V}_J$ — components' parameters

$\boldsymbol{\pi} = \pi_1, \ldots, \pi_J; \sum_j \pi_j = 1$ mixing weights (probabilities) for each component

$\boldsymbol{C} = C_1, \ldots, C_J$ — J classes C_j

What is Finite Mixture Model? (cont.)

Basic Model

$$p(X_i|\boldsymbol{V},\boldsymbol{\pi},J) = \sum_{j=1}^{J} \pi_j \cdot p(X_i|X_i \in C_j,\boldsymbol{V})$$ -- **Instance**

Conditionally Independent Instances Assumption

$$p(\boldsymbol{X}|\boldsymbol{V},\boldsymbol{\pi},J) = \prod_{i=1}^{I} p(X_i|\boldsymbol{V},\boldsymbol{\pi},J)$$ -- **Database**

[11][12]

Applying Bayes's Theorem

$$p\left(X_i \in C_j \middle| X_i,\boldsymbol{V},\boldsymbol{\pi},J\right) = \frac{\pi_j \cdot p(X_i|X_i \in C_j,\boldsymbol{V})}{p(X_i|\boldsymbol{V},\boldsymbol{\pi},J)}$$ -- **Classification**

What to learn?

Problem

Identify a finite mixture.

Two Parts

1. Determine the classification **parameters** for a given number of classes

$$p(\boldsymbol{V},\boldsymbol{\pi}|\boldsymbol{X},J)=\frac{p(\boldsymbol{V},\boldsymbol{\pi}|J)\cdot p(\boldsymbol{X}|\boldsymbol{V},\boldsymbol{\pi},J)}{p(\boldsymbol{X}|J)}$$

Prior distribution of the parameters

Likelihood function

Pseudo-likelihood as a constant

[11][12]

2. Determine the **number** of classes for a given database

$$p(J|\boldsymbol{X})=\frac{p(J)\cdot p(\boldsymbol{X}|J)}{p(\boldsymbol{X})}$$

Prior probability of the number of classes

Prior probability of the database

How to learn?

- **What do we have?**

  Unlabeled data with **unknown distributions**.

  Latent data **Z** determining classes

  Unknown parameters θ including mixture & components

- **Maximum Likelihood Estimate (MLE)**

  Likelihood function: $L(\theta; \boldsymbol{X}, \boldsymbol{Z}) = p(\boldsymbol{X}, \boldsymbol{Z}|\theta)$ [13]

  Marginalization: $L(\theta; \boldsymbol{X}) = p(\boldsymbol{X}|\theta) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\theta)$

  Does it work?

- **If NOT, then what?**

How to learn? (cont.)

- **Expectation–maximization (EM)**

  Iteratively doing two steps until convergence to estimate parameters.

- **Algorithm**

1. Initialization

   $\theta_0 \in \Theta$

2. Iteration

   For t=0,1,2…until convergence

   ➢ Expectation step (E step)

   $$Q(\theta, \theta_t) = E_{\theta_t}(log p_\theta(\boldsymbol{Z}, \boldsymbol{X})|\boldsymbol{X}) = \sum_Z p_{\theta_t}(Z|X) \cdot \log p_\theta(Z, X)$$

   ➢ Maximization step (M step)

   $\theta_{t+1} \in argmax_\theta Q(\theta, \theta_t)$

Conditional Expectation

"Complete" data

[14][15][16]

How to learn? (cont.)

- **EM Example**

Data set with two binary attributes A1 and A2:

	A_1	A_2
D1	1	1
D2	0	0

Initial parameter values:

$$P(z = 0) = P(z = 1) = 0.5$$

	$A_1 = 0$	$A_1 = 1$
$P(A_1 \mid z = 0)$	0.6	0.4
$P(A_1 \mid z = 1)$	0.4	0.6

	$A_2 = 0$	$A_2 = 1$
$P(A_2 \mid z = 0)$	0.6	0.4
$P(A_2 \mid z = 1)$	0.4	0.6

How to learn? (cont.)

- **EM Example**

E-Step:

$$P(z=0|D1) = \frac{0.5 * 0.4 * 0.4}{0.5 * 0.4 * 0.4 + 0.5 * 0.6 * 0.6} = 0.31$$

$$P(z=1|D1) = 0.69; P(z=0|D2) = 0.69; P(z=1|D2) = 0.31$$

M-Step:

The parameter values after first iterations

$$P(z=0) = P(z=1) = 0.5$$

	$A_1 = 0$	$A_1 = 1$
$P(A_1\|z=0)$	0.69	0.31
$P(A_1\|z=1)$	0.31	0.69

	$A_2 = 0$	$A_2 = 1$
$P(A_2\|z=0)$	0.69	0.31
$P(A_2\|z=1)$	0.31	0.69

How to learn? (cont.)

- **Life is not that fancy!**

  What do we want to know?

  Parameters for the mixture: J, $\boldsymbol{\pi}$

  Parameters for the components: $\boldsymbol{V}$

  But…

  Suppose J is known,

  if we want to know how many instances in each class (which is of the same essential meaning as $\boldsymbol{\pi}$), we have to know how each class describes its instances;

  if we want to know how each class is described ($\boldsymbol{V}$), we have to know how instances settle in these classes.

  What are we going to do?

How to learn? (cont.)

- **Conjugate prior distribution for $\boldsymbol{\pi}$**

  Class is a categorical distribution.

  Conjugate wrt Dirichlet distribution

Make it Bayesian!

Categorical distribution
vs
Multinominal distribution

Analog

Bernoulli distribution
vs
Binominal distribution

- **Estimation System**

# of conjugate prior fictitious data points

$$\hat{\pi}_j = \frac{W_j + w' - 1}{I + J(w' - 1)}$$

$$\frac{\partial}{\partial V_j} \ln p(\hat{V}_j) + \sum_{i=1}^{I} w_{ij} \frac{\partial}{\partial V_j} \ln p(X_i | \hat{V}_j) = 0$$

Treat w_{ij} as constant

$$w_{ij} = p(X_i \in C_j | X_i, \hat{V}_j, \hat{\pi}_j)$$

$$W_j = \sum_{i=1}^{I} w_{ij}$$

[11][17]

Treat $\boldsymbol{V}, \boldsymbol{\pi}$ as constants

$$p(X_i \in C_j | X_i, \boldsymbol{V}, \boldsymbol{\pi}, J) = \frac{\pi_j \cdot p(X_i | X_i \in C_j, \boldsymbol{V})}{p(X_i | \boldsymbol{V}, \boldsymbol{\pi}, J)}$$

How to learn? (cont.)

- **How about J ?**

  What are the extreme cases?

  1 and *I*.

  Where is the answer?

  Somewhere in between.

  How to find it?

  Search (local search in most of the times).

  How the searching result would be?

  Completeness? Yes. Optimality? No.

  How to search?

How to learn? (cont.)

- **Searching for J**

  Try to imagine you are sprinkling something…

  And some others are catching…

- **Essence**

  Multi-restart for breaking local maxima.

Rules

- Start with numerous catchers there, more than we are expecting.
- Who can always catch a lot stay, otherwise out!

AutoClass Attribute Models

- AutoClass provides basic models for several types of numerical data.
- All class models share the same attribute set
- In each case they adopt a minimum information prior

Attribute model for Various types of data

Discrete valued attributes:

Bernoulli distributions

$$\mathsf{P}(X_{ik} = l \mid X_i \in C_j, |\vec{V}_{jk}, T_{jk}, S, \mathcal{I}) \equiv q_{jkl}$$

Prior: dirichlet conjugate

$$\mathsf{P}(q_{jk1}, \ldots, q_{jkL_k} \mid T_{jk}, S, \mathcal{I}) \equiv \frac{\Gamma(L_k + 1)}{[\Gamma(1 + \frac{1}{L_k})]^{L_k}} \prod_{l=1}^{L_k} q_{jkl}^{\frac{1}{L_k}}$$

$$\widehat{q}_{jkl} = \frac{w_{jkl} + \frac{1}{L_k}}{w_j + 1}$$

For covariant case, they apply the model to the cross product of individual attribute values

e.g: Female and blood type A

Attribute model for Various types of data

Real valued location attributes:

$$\mathsf{P}(X_{ik} \mid X_i \in C_j, \mu_{jk}, \sigma_{jk}, T_{jk}, S, \mathcal{I}) \equiv \frac{1}{\sqrt{2\pi}\sigma_{jk}} e^{-\frac{1}{2}\left(\frac{x_{ik}-\mu_{jk}}{\sigma_{jk}}\right)^2}$$

$$\mathsf{P}(\mu_{jk} \mid T_{jk}, S, \mathcal{I}) = \frac{1}{\mu_{k_{max}} - \mu_{k_{min}}}, \qquad \widehat{\mu}_{jk} = m_{jk}$$

$$\mathsf{P}(\sigma_{jk} \mid T_{jk}, S, \mathcal{I}) = \sigma_{jk}^{-1} \left[\log \frac{\sigma_{k_{max}}}{\sigma_{k_{min}}}\right]^{-1}, \qquad \widehat{\sigma}_{jk}^2 = s_{jk}^2 \frac{w_j}{w_j + 1}.$$

Uniform prior for means

Jeffreys prior for singleton attribute's standard deviation

Missing data

‘Unknown’ has to be treated as a valid data value. Discarding the unknown value may destroys potentially valuable information.

Discrete value: add ‘missing’ as an additional value

Numerical attributes: use a binary discrete probability for ‘missing’

$$\mathrm{P}(X_{ik} = missing \mid X_i \in C_j, q_{jk}, \mu_{jk}, \sigma_{jk}, T_{jk}, S, \mathcal{I}) \equiv q_{jk},$$

$$\mathrm{P}(X_{ik} = r \mid X_i \in C_j, q_{jk}, \mu_{jk}, \sigma_{jk}, T_{jk}, S, \mathcal{I}) \equiv \frac{(1 - q_{jk})}{\sqrt{2\pi}\sigma_{jk}} e^{-\frac{1}{2}\left(\frac{r - \mu_{jk}}{\sigma_{jk}}\right)^2}.$$

Hierarchical models

- To deal with duplication problem, e.g. ‘dog’ and ‘cat’
- Represent the mixture model in tree structure where multiple classes can share one or more model terms

Hanson, Robin, John Stutz, and Peter Cheeseman. "Bayesian Classification with Correlation and Inheritance." *IJCAI*. 1991.

Irrelevant attributes

- To eliminate the attribute which is deemed irrelevant to only some of models will cause error

e.g. Given two mode $\vec{V}_j, T_j$ and $\vec{V}_j', T_j'$ identical in both form and parameter values except that the latter modeling additional attribute. Then for any instance:

$$\mathrm{P}(\vec{X}_i \mid X_i \in C_j, \vec{V}_j, T_j, S, \mathcal{I}) \;>\; \mathrm{P}(\vec{X}_i \mid X_i \in C_j', \vec{V}_j', T_j', S, \mathcal{I}).$$

Deal with irrelevant attributes:

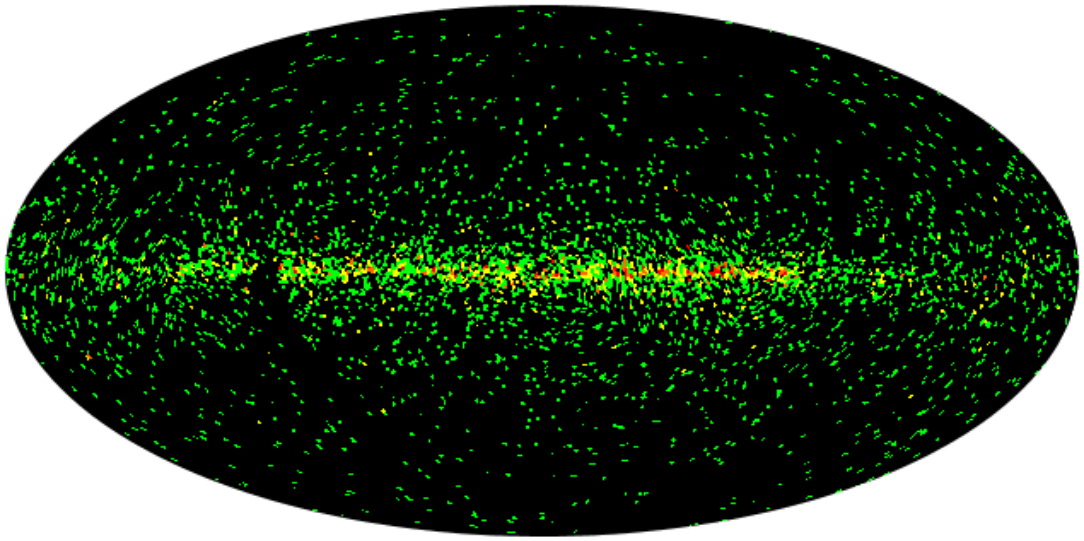
- An attribute is irrelevant when all classes possess identical p.d.f.'s for that attribute
- In the hierarchical model, push the attribute model up to the root node

This will allow comparison between classifications that deem different attribute subset irrelevant

Infrared Astronomical Satellite (IRAS) Case Study

IRAS Low Resolution Spectral Atlas:

- 5425 mean spectra of IRAS points sources.
- For each spectrum, there are 100 blue channels and 100 red channels and 100 of the 200 channels are usable.
- Treat each of the 100 spectral channels as an independent normally distributed single real value

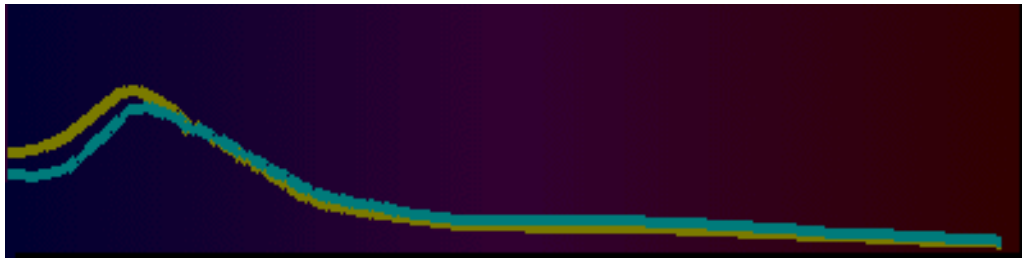


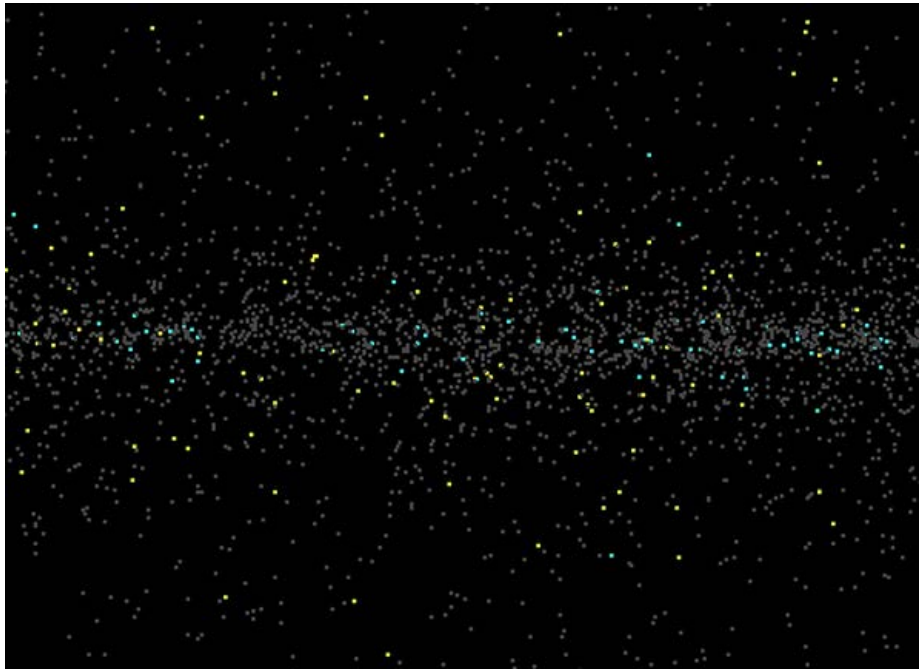
Galactic Ecliptic Equatorial

How to learn? (cont.)

Observation:

- AutoClass classes are significantly different from that provided with the atlas
- AutoClass is able to make distinctions between spectra that looks similar





http://ti.arc.nasa.gov/tech/rse/synthesis-projects-applications/autoclass/

DNA Intron Data

Exon: the segment of DNA that contribute to the final RNA

Donor site: At the beginning of exon/intron boundary

Acceptor site: At the end of exon/intron boundary

Database: 3000 donor and acceptor sites from human DNA

Assumption: in human DNA there is only one general type of donor and acceptor site since they all use the same splicing machinery.

DNA Intron Data

3 classes with obvious pattern:

largest class: C rich, every position has a significantly higher probability of having a C than the global average

The other 2 classes are TA rich and G rich

Question: whether the class of donor site is correlated with the class of the corresponding acceptor site

- The class of a donor site is highly correlated with the corresponding acceptor site
- The same classes were observed in mouse genes
- The base-frequency pattern extends into the flanking exons, but not as strongly as that observed in the introns
- If one intron is TA rich, there is high probability that any neighboring introns will also be TA rich

LandSat Data

http://landsat.usgs.gov/

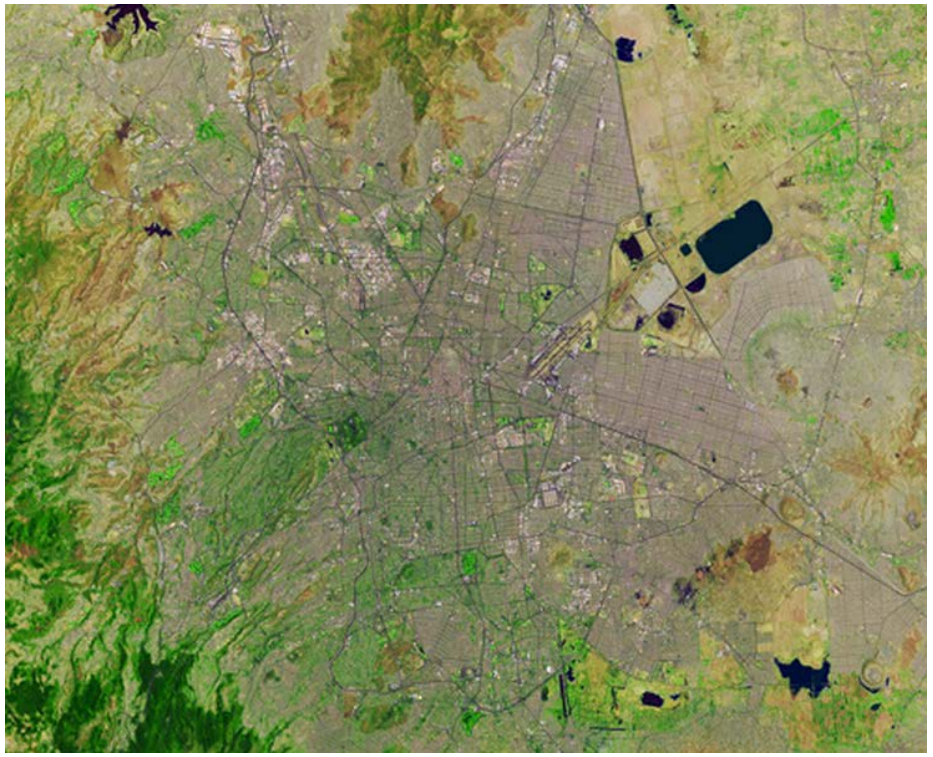
Database: A 1024*1024 array of LandSat pixels

LandSat pixel: record 7 spectral intensities from a 30m square ground patch

The goal is to find classes in the set of over 1 million pixels

The correlation between values within each class were considered. (Hanson, 1991)

They used the ratio of spectral intensities for each pixel instead of the ground slope of patch.



LandSat 8 - Mexico City

LandSat Data

Results:

- classification found 93 classes, meta-classification, which made the individual classes easier to interpret.
- Assign each pixel to its most probable class, then plot the resulting classes. For many classes, these results suggest an interpretation to the human eye (e.g. roads rivers, valley bottoms and valley edges).
- Pixels with a mixture of basic types: the classes with mixture of basic type may not be particularly meaningful. Majority of classes seems to be composed of pure pixels of a single type.

What is more…

- **Parametric or Non-Parametric ?**

  On a size n sample space, apply Kernel Estimation.

  If g=n, Non-parametric. $\hat{f}(y_j) = \frac{1}{nh}\sum_{i=1}^{n} k((y_j - y_i)/h)$ [2]

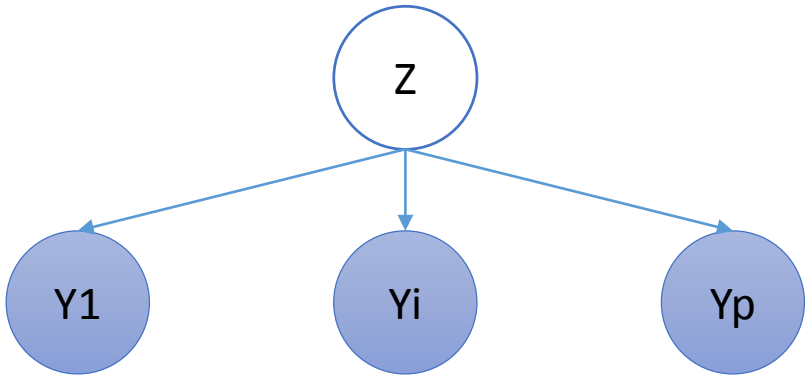
  if g=1, fully parametric with a single parametric family.

  Trade-off between bias and variance

- **Independent Assumption on Y_i**

  Similar as Naive Bayes Model

  $$\Pr(\boldsymbol{y}) = \Pr(y_1, \ldots, y_p) = \sum_{k=1}^{g} \Pr(Z = k) \prod_{i=1}^{p} \Pr(y_i | Z = k)$$



What is more…

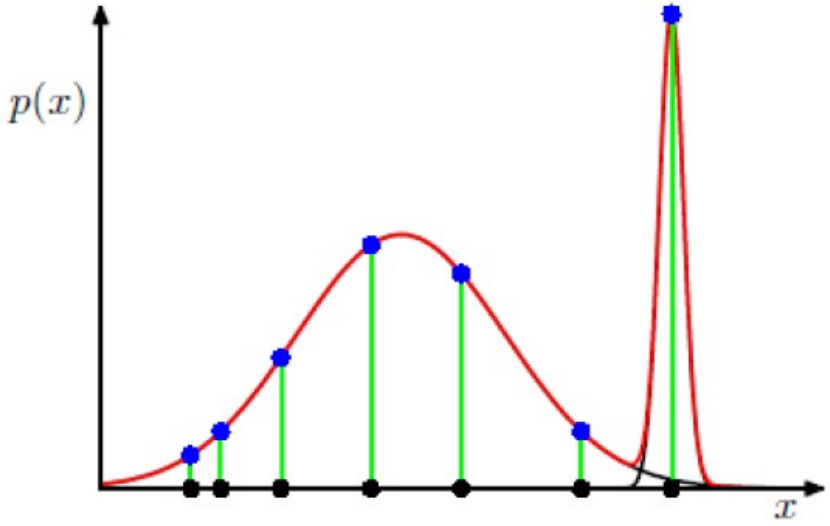
- **What if T_i are different?**

  We cannot always assume T_i are all of Normal Distribution. [11]

- **The singularity problem in EM?**

  Causes: outliers and repeated points.

  Solution: prior distribution, multi-start,…



[5]

References

[1] Partha Deb, "*Finite Mixture Models*", http://www.stata.com/meeting/snasug08/deb_fmm_slides.pdf , July 2008

[2] Geoffrey McLachlan and David Peel, "*Finite Mixture Models*", October 2000

[3] Mario A.T. Figueiredo and Anil K. Jain, "*Unsupervised Learning of Finite Mixture Models*", March 2002

[4] Karl Pearson, "*Contributions to the Mathematical Theory of Evolution*", January 1894

[5] Nevin L. Zhang, "Introduction to Bayesian Networks", http://www.cse.ust.hk/~lzhang/teach/5213/slides/l10.p.pdf

[6] Dave Kessler and Allen McDowell, "*Introducing the FMM Procedure for Finite Mixture Models*", March 2012

[7] Jeff A. Bilmes, "*Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*", April 1998

[8] Douglas Reynolds, "*Gaussian Mixture Models*", 2009

[9] Fabio Gagliardi Cozman, "*Semi-Supervised Learning of Mixture Models and Bayesian Networks*", March 2003

[10] Carey E. PRIEBE, "*Adaptive Mixtures*", September 1994

References

[11] Peter Cheeseman, John Stutz, *"Bayesian classification (AutoClass): theory and results"*, 1996

[12] Cheeseman, Peter; Kelly, James; Self, Matthew; Stutz, John; Taylor, Will; Freeman, Don, *"AutoClass: A Bayesian classification system"*, 1988

[13] Wikipedia, *"Maximum likelihood"*, http://en.wikipedia.org/wiki/Maximum_likelihood_estimate

[14] Wikipedia, *"Expectation–maximization algorithm"*, http://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

[15] Eugene Weinstein, *"Expectation-Maximization Algorithm and Applications"*, http://cs.nyu.edu/~eugenew/publications/em-talk.pdf, November 2006

[16] Jeff A. Bilmes, *"A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models"*, 1998

[17] Hagit Shatkay, *"Computational Biomedicine Notes"*, March 2014

[18] Morris H. DeGroot, "Probability and Statistics (4th Edition)"

[19] Geoffrey J McLachlan, "The EM algorithm and extensions", 1997

References

[20] Daphne Koller, “Probabilistic Graphical Models: Principles and Techniques”, 2009