Probabilistic classification models

Peter McCullagh

Department of Statistics
University of Chicago

Summer school on machine learning
University of Chicago
June 3 2009

Based on joint work with Jie Yang

www.stat.uchicago.edu/~pmcc/seminars/discrimination.pdf
Outline

**Statistical classification**
- Random partitions
- Ewens partition process

**Processes**
- The Gauss-Ewens process
- Illustrations
- Classification distribution
- Illustration

**A point process model**
- Cox processes
- Probability distributions for point configurations
- Boson point processes
- Boson classification distributions
- Illustration

**References**
Statistical classification and discrimination

Given training data on \( n \) units (specimens) consisting of
Feature values \( Y = (Y_1, \ldots, Y_n) \) \( Y_i \in \mathcal{R}^d \)
A partition \( B \) of the specimens into blocks or classes \( \{b\} \)

Given a new unit \( u' \) with feature value \( Y(u') \)
compute the class probabilities \( \text{pr}(u' \mapsto b \mid \text{data}) \)

The classes are either a fixed labelled set \( \{0, 1, \ldots, 9\} \)
\( \ldots \) or the blocks \( b \in B \)
\( \ldots \) or the blocks of \( B \) plus the empty set

Need something akin to an exchangeable process
May need to estimate some parameters from \( (Y, B) \)
# Partitions

$[n] = \{1, \ldots, n\}$ a finite set of specimens

A partition $B$ of the set $[n] = [6]$ is:

(i) a set of disjoint non-empty subsets $b \subset [n]$ called blocks...
   
   e.g. $B = \{\{2, 4, 6\}, \{1, 3\}, \{5\}\} \equiv 246|13|5 \equiv 13|246|5$

(ii) an equivalence relation $B: [n] \times [n] \rightarrow \{0, 1\}$
   
   s.t. $B(i, j) = 1$ if $i \sim j$ belong to the same block

(iii) a symmetric Boolean matrix

\[
B = \begin{pmatrix}
1 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 \\
1 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 \\
\end{pmatrix}
\]

$\#B$: number of elements (no. of blocks)

for $b \in B$, $\#b$ is the number of elements ($\#b > 0$)

Integer partition $\nu(B) = 1 + 2 + 3$ associated with $B = 246|13|5$
The set $\mathcal{E}_n$ of partitions of $[n]$

$\mathcal{E}_1$: 1
$\mathcal{E}_2$: 12, 1|2
$\mathcal{E}_3$: 123, 12|3, 13|2, 23|1, 1|2|3
$\mathcal{E}_4$: 1234, 123|4[4], 1234[3], 123|4[6], 1|2|3|4
$\mathcal{E}_5$: 12345, 1234|5[5], 12345[10], 1234|5[10], 123|45[15], 123|4|5[10], 1|2|3|4|5

$\#\mathcal{E}_n$: 1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975, 678570

Permutation map $\pi : [n] \rightarrow [n]$ also acts $\pi^* : \mathcal{E}_n \rightarrow \mathcal{E}_n$

Partition type $\nu(B)$ is maximal invariant

Deletion map: $D_n : \mathcal{E}_n \rightarrow \mathcal{E}_{n-1}$ (onto)

$D_4$: 1234, 123|4 $\leftrightarrow$ 123

12|3|4, 12|34, 124|3 $\leftrightarrow$ 12|3 (three types)

1|2|3|4, 1|2|34, 1|24|3, 14|2|3 $\leftrightarrow$ 1|2|3

Same as removal of last row and column from matrix

$\mathcal{E}$ represents the sets $\{\mathcal{E}_n\}$ with permutation and deletion maps
Probability distributions on partitions

$P_n$ a probability distribution on $\mathcal{E}_n$
Finitely exchangeable if $\nu(B) = \nu(B')$ implies $P_n(B) = P_n(B')$
Examples:

| $\mathcal{E}_3$ | 123 | 12|3 | 13|2 | 23|1 | 1|2|3 |
|---|---|---|---|---|---|---|
| $P_3$ | 1/3 | 1/6 | 1/6 | 1/6 | 1/6 |
| $P'_3$ | 1/6 | 1/6 | 1/6 | 1/6 | 1/3 |

| $\mathcal{E}_4$ | 1234 | 123|4 | 13|24 | 23|1|4 | 1|2|3|4 |
|---|---|---|---|---|---|---|---|
| $P_4$ | 1/4 | 1/12 | 1/24 | 1/24 | 1/24 |
| $P'_4$ | 1/10 | 1/15 | 1/30 | 1/15 | 2/15 |

Compatibility:

$P_4(1234 \cup 123|4) = 1/4 + 1/12 = 1/3 = P_3(123)$
$P_4(12|3|4 \cup 12|34 \cup 124|3) = 2/24 + 1/12 = 1/6 = P_3(12|3)$
$P_4(1|2|3|4 \cup 1|2|34[3]) = 1/24 + 3/24 = 1/6 = P_3(1|2|3)$

$P_3$ is the marginal distribution of $P_4$ under deletion

$P'_3$ is the marginal distribution of $P'_4$ under deletion
Exchangeable partition process

An exchangeable partition process is a sequence $P_n$ such that each $P_n$ is finitely exchangeable. $P_n(B)$ depends only on block sizes $\nu(B)$. $P_n$ is the marginal distribution of $P_{n+1}$.

Kolmogorov compatibility condition:

$$P_n(B) = \sum_{B' : D_{n+1} B' = B} P_{n+1}(B')$$

Conditional distribution

$$P_{n+1}(B' | B \in \mathcal{E}_n) = \begin{cases} P_{n+1}(B')/P_n(B) & D_{n+1} B' = B \\ 0 & \text{otherwise.} \end{cases}$$

Kingman’s paintbox characterization
The Ewens partition process

Ewens distribution with parameter $\lambda > 0$

$$P_n(B) = \frac{\Gamma(\lambda) \lambda^{\#B}}{\Gamma(n + \lambda)} \prod_{b \in B} \Gamma(\#b) \quad (B \in \mathcal{E}_n)$$

Conditional distributions

$$P_{n+1}(u_{n+1} \mapsto b \mid B) = \frac{P_{n+1}(B')} {P_n(B)} = \begin{cases} \frac{\#b}{n + \lambda} & b \in B \\ \frac{\lambda}{n + \lambda} & b = \emptyset \end{cases}$$

(Pitman’s CRP description)

Induced distribution on integer partitions $\nu(B) = 1^{\nu_1} 2^{\nu_2} \cdots n^{\nu_n}$

$$Q_n(\nu) = \frac{\Gamma(\lambda) \lambda^\nu}{\Gamma(n + \lambda)} \prod_{j=1}^{n} ((j - 1)!)^{\nu_j} \times \frac{n!}{\prod (j!)^{\nu_j} \nu_j!}$$

No deletion operation for integer partitions
Hence no process on integer partitions
Other interpretations of the Ewens process

Conditional Poisson interpretation
$X_1, X_2, \ldots$ independent Poisson variables $X_j \sim \text{Po}(\lambda/j)$ as multiplicities $1^{X_1} 2^{X_2} 3^{X_3} \ldots$ in integer partition

Conditional distribution of $X_1, \ldots, X_n$ given $\sum jX_j = n$

$$p(x) \propto e^{-\lambda \sum 1/j} \frac{\lambda^x}{\prod j^{x_j} x_j!}$$

is exactly the Ewens partition

Negative binomial model for the number of species
Fisher 1940; Good 1953; Mosteller & Wallace 1964; Efron & Thisted 1976
Kendall’s (1975) family-size process (Kelly, 1979)
Prime factorization (Billingsley 1972; Donnelly & Grimmett)
Partition induced by Dirichlet process (Jordan,...)
Characterization of the Ewens distribution

Why is the Ewens distribution ubiquitous?

(i) Exchangeability: \( B \sim P_n \) implies \( B^\pi = \pi^{-1} B^\pi \sim P_n \)

(ii) Restriction to subsets \([m] \subset [n]\)

\[ \text{if } B \sim P_n, \text{ the restriction is } B[m] \sim P_m \text{ (process property)} \]

(iii) Self-similarity (lack of memory)

Given that \( B \leq b|b' \), conditional distn is \( B \sim P_{\#b} \times P_{\#b'} \)

(Aldous, 1994)

Leading to a theory of Markov fragmentation trees...

by recursive partitioning...
The Gauss-Ewens cluster process

Components of a cluster process:
(i) An index set $\mathbb{N}$ (specimens, plots, patients, units,)
(ii) A random sequence $Y_1, Y_2, \ldots$ with $Y_i \in S$, ($S = \mathbb{R}^d$)
(iii) A random partition $B$ of $\mathbb{N}$ (not a partition of $S$!)
(iv) For each sample $[n] \subset \mathbb{N}$ a prob distribution $P_n$

Gauss-Ewens distribution $P_n$ on $S^n \times \mathcal{E}_n$

$$P_n(B[n] = B) = \frac{\lambda^{\#B} \Gamma(\lambda)}{\Gamma(n + \lambda)} \prod_{b \in B} \Gamma(\#b)$$

$$Y[n] \mid B[\mathbb{N}] \sim N(1_\mu, l_n \otimes \Sigma + B[n] \otimes \Sigma')$$

Note difference between $B[n]$ and $B[\mathbb{N}]$ (no interference)

Parameters: $\mu \in \mathbb{R}^d$, $\lambda > 0$;
$\Sigma, \Sigma'$ within- and between-cluster covariance matrices
Exchangeability of cluster process

Observation for a finite sample \([n] \subset \mathbb{N}\) is \((Y[n], B[n])\)

Observation space is \(S^n \times \mathcal{E}_n\)

Permutation \(\pi : [n] \rightarrow [n]\) acts on observation space
\((Y[n], B[n]) \mapsto (Y^\pi, B^\pi)\) by composition \(B^\pi(i, j) = B(\pi_i, \pi_j)\)

Restriction \(\varphi : [m] \rightarrow [n]\) acts on observation spaces
\((Y[n], B[n]) \mapsto (Y^\varphi, B^\varphi)\) by composition
\(Y^\varphi(i) = Y(\varphi(i)), \quad B^\varphi(i, j) = B(\varphi_i, \varphi_j)\)

(i) Distribution \(Q_n\) on \(\mathcal{E}_n \times S^n\) unaffected by permutation \(\pi\) of \([n]\)
(ii) \(Q_m\) on \(\mathcal{E}_m \times S^m\) is the marginal distn of \(Q_n\)

Hence there exists an infinite random clustering \((Y, B)\)
Constructive version of Gauss-Ewens process

Four-stage construction of Gauss-Ewens process

\[ \eta \sim IIDN(\mu, \Sigma'), \quad \text{cluster centres in } \mathcal{R}^d \]
\[ \epsilon \sim IIDN(0, \Sigma), \quad \text{independent of } \eta \]
\[ tbl_{i+1} = \text{one of } (tbl_1, \ldots, tbl_i) \text{ with equal prob } 1/(i + \lambda) \]
\[ \text{else new w. p. } \lambda/(i + \lambda) \]
\[ Y_i = \eta_{tbl(i)} + \epsilon_i \]

with \( (Y_1, tbl_1), \ldots, (Y_n, tbl_n) \) observed in training
Parameters \( \lambda > 0, \mu \in \mathcal{R}^d, \Sigma, \Sigma' \) of order \( d \)

For more esoteric structures, include cluster shape info in \( \eta \)
\[ Y_i = g(\eta_{tbl(i)}, \epsilon_i) \]
Ordinary Gauss-Ewens process in $\mathcal{R}^2$
Gauss-Ewens process with sub-clusters in $\mathcal{R}^2$
Gauss-Ewens process with topological clusters
Cluster models for classification w/o classes

Problem: No set $\mathcal{C}$ of classes in a cluster process $(Y, B)$
Observation $(Y, B)[n]$ in training sample $u_1, \ldots, u_n$
How can we assign new unit $u_{n+1}$ to classes?

Conditional distribution

$$\Pr(u_{n+1} \mapsto b \mid (Y, B)[n], y') = \begin{cases} f(\ldots) & b \in B \\ \ldots & \text{otherwise.} \end{cases}$$

Blocks of $B$ are the classes!
Also need parameter estimates (at least $\lambda, \theta = \Sigma' \Sigma^{-1}$)

Lack of $\mathcal{C}$ is not a disadvantage!
possibility of assigning $u_{n+1}$ to a previously unseen class
Explicit calculation of conditional distribution

Simplification $\Sigma' = \theta \Sigma$ in $S = \mathcal{R}^d$

$$\text{pr}(u' \mapsto b \mid y', ...) \propto \begin{cases} \# b \phi_d(y' - \tilde{\mu}_b; \tilde{\Sigma}_b) & b \in B \\ \lambda \phi_d(y(u') - \mu; \Sigma(1 + \theta)) & b = \emptyset \end{cases}$$

$$\tilde{\mu}_b = \frac{\mu + n_b \theta \bar{y}_b}{1 + n_b \theta}, \quad \tilde{\Sigma}_b = \Sigma(1 + \theta/(1 + n_b \theta))$$

Typical values $\theta \geq 5$ and $n_b \geq 5$

so $|\tilde{\mu}_b - \bar{y}_b| \leq 0.04|\mu - \bar{y}_b|$

(similar to Fisher discriminant model, but with shrinkage)

Tree version with classes and sub-classes
Block having maximum conditional probability
A point process model for classification

\( \mathcal{C} \): a finite set of classes, \{0, 1, \ldots, 9\}

\( S \): the feature space, usually \( \mathcal{R}^d \)

\( \lambda \): a random intensity function on \( \mathcal{C} \times S \)

\( \Lambda \) is a \( \sigma \)-finite random measure with density \( \lambda \)

Given \( \Lambda \), \( Z \subset \mathcal{C} \times S \) is a Poisson process: \( Z \sim \text{Po}(\Lambda) \)

Each event \( z \in Z \) is a pair \( z = (x, y) \)

- class of event \( x(z) \); feature of event \( y(z) \)

Observe \( Z = \{(x_1, y_1), \ldots, (x_n, y_n)\} \subset S \) (finite in finite time)

Next event \( z' \) has feature \( y(z') \) observed

What is the conditional distribution of \( x(z') \) given the data?

Not an exchangeable process because

(i) no index set of specimens

(ii) rate of occurrence of events is random
Point process distributions: outline for general $\mathcal{X}$

Random measure $\Lambda$ non-atomic in domain $\mathcal{X}$
$X \subset \mathcal{X}$ is conditionally Poisson($\Lambda$)
Test set $S \subset \mathcal{X}$ such that $\Lambda(S) < \infty$ w.p.1
Observation $x = X \cap S$ has density in configuration space

$$p_n(x) = E \left( \lambda(x_1) \cdots \lambda(x_n) \exp(-\Lambda(S)) \right)$$

$$\sum_{n=0}^{\infty} \frac{1}{n!} \int_{S^n} p_n(x) \, dx = 1$$

Papangelou conditional intensity at $x' \in S$

$$E(\lambda(x') \mid X \cap S = x) = \frac{p_{n+1}(x \cup x')}{p_n(x)}$$

is also the Bayes estimate of the intensity.
$\alpha$-permanent: definition and properties

$$\text{per}_\alpha(A) = \sum_\sigma \alpha^{\#\sigma} \prod_{j=1}^n A_{j,\sigma_j}$$

$$\text{cyp}(A) = \sum_{\sigma: \#\sigma=1} \prod_{j=1}^n A_{j,\sigma_j} = \lim_{\alpha \to 0} \alpha^{-1} \text{per}_\alpha(A)$$

$\#\sigma$ is the number of cycles

$\text{per}_\alpha(1_{n\times n}) = \alpha^\up{n} = \Gamma(n + \alpha) / \Gamma(\alpha)$

Convolution property: $(K[x])_{i,j} = \{K(x_i, x_j)\}$ of order $\#x$

$$\sum_{w \subseteq x} \text{per}_\alpha(K[w]) \text{per}_{\alpha'}(K[w]) = \text{per}_{\alpha+\alpha'}(K[x])$$
Boson point process

$K(x, x')$: a positive definite function $\mathcal{X} \times \mathcal{X} \to \mathcal{R}$, symm and cts
$W$: a Gaussian process in $\mathcal{X}$; covariance function $K/2$
$W_1, \ldots, W_k$: iid copies of $W$
$\lambda(x) = W_1^2(x) + \cdots + W_k(x)^2$: a random intensity in $\mathcal{X}$

Poisson point process generated by $\lambda$ has cumulant density

$$\text{cum}(\lambda(x_1), \ldots, \lambda(x_n)) = \alpha \text{cyp}(K[x])$$

moment density

$$E(\lambda(x_1) \cdots \lambda(x_n)) = \text{per}_\alpha(K[x])$$

and probability density

$$p_n(x) = \text{const} \times \text{per}_\alpha(\tilde{K}[x])$$

$\alpha = k/2$, $\tilde{K} = K(I + K)^{-1}$
Boson multi-class model

Parameter $\alpha_r$ associated with class $r$
Intensity for class $r$, $\lambda_r(x) = W^2_1(x) + \cdots + W^2_{2\alpha_r}(x)$
  independent intensity for each class
Superposition intensity $\lambda(x) = W^2_1(x) + \cdots + W^2_{2\alpha}(x)$
Observation $(X, Y) \sim (y^{(1)}, \ldots, y^{(k)})$
  $y^{(r)}$ feature values for class $r$: $y^{(r)} \perp y^{(s)}$ (indep)
  superposition of features $y = \bigcup y^{(r)}$

Joint distribution is a product of Boson processes

$$p(y^{(1)}, \ldots, y^{(k)}) = \text{const} \times \text{per}_{\alpha_1}(\tilde{K}[y^{(1)}]) \cdots \text{per}_{\alpha_k}(\tilde{K}[y^{(k)}])$$

$$p(y) = \text{const} \times \text{per}_{\alpha}(\tilde{K}[y])$$

leading to a multinomial-type conditional distribution $p(X \mid Y)$ on labelled partitions of $y$. 
Classification distributions: labelled and unlabelled

Labelled partition \((X, Y)\)
Features observed in training: \((X, Y) \equiv (\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(k)})\)
(Some classes may be empty \(\mathbf{y}^{(r)} = \emptyset\).

\[
\Pr(u' \mapsto r \mid (X, Y), y') \propto \frac{\text{per}_{\alpha_r}(\tilde{K}[\mathbf{y}^{(r)}, y'])}{\text{per}_{\alpha_r}(\tilde{K}[\mathbf{y}^{(r)}])} \quad 1 \leq r \leq k
\]

For the unlabelled partition \((B, Y)\), \(#B \leq k\) and \(\mathbf{y}^{(b)} \neq \emptyset\)
Limit for \(\alpha_1 = \cdots = \alpha_k = \lambda/k\) and \(k \to \infty\) with \(\lambda\) fixed

\[
\Pr(u' \mapsto b \mid (B, Y), y') \propto \begin{cases} 
\text{cyp}(\tilde{K}[\mathbf{y}^{(b)}, y'])/\text{cyp}(\tilde{K}[\mathbf{y}^{(b)}]) & b \in B \\
\lambda \tilde{K}(y', y') & b = \emptyset
\end{cases}
\]

with \(X\) replaced by unlabelled partition \(B\)
Example: $\alpha_1 = \alpha_2; \quad \tilde{K}(y, y') = \exp(-|y - y'|^2/\tau^2)$
Density plot of predictive probability $\text{pr}(\text{red} \mid \text{data})$
Algorithms for approximation

Exact algorithms for $\alpha$-permanents:
- Ryser algorithm OK for $n \leq 20$, general $K$ and $\alpha = 1$
- Barvinok’s algorithm OK for $\alpha = 1$, low rank $K$ and $n \leq 30$

Stochastic approximation algorithms
- Karmarkar’s algorithm: nice idea for $\alpha = 1$, but ...
- Importance-sampling algorithm (Kou and McC)

Markov chain algorithms (non-negative $K$, $\alpha = 1$)
- Jerrum, Sinclair and Vigoda: demonstration of principle

For $\alpha$-permanent ratio:
- Truncated cycle expansion (McC and Yang)
  ... naive, but it seems to work for general $K$, $\alpha$
- Importance-sampling algorithm (Kou and McC)
  achieves good accuracy fairly quickly for general $K$, $\alpha$
References

Kelly, F. (1979) Reversibility and Stochastic Networks. Wiley