Classification

Allan Aasbjerg Nielsen
alan@dtu.dk

Technical University of Denmark
DTU Compute – Applied Mathematics and Computer Science

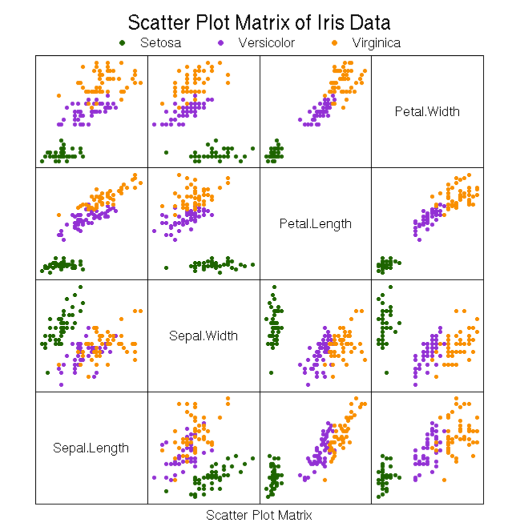
3 Feb 2017

Classification

- classification is the process of grouping observations (pixels or regions) into classes intended to represent different physical objects or types
- here, the production of a **thematic map** from (image) data with digital numbers representing for example reflected or emitted EM-radiation in different wavelength bands
- very many classification methods ranging from quite simple to highly advanced
- two major groups of methods: supervised and unsupervised
  - supervised: ideally physical classes but not necessarily statistically distinct
  - unsupervised: statistically distinct but not necessarily physical classes

Feature space

1. p variables
   C classes
   N observations (or samples)
2. $\boldsymbol{x}_i,\ i = 1, \ldots, N,\ p \times 1$
   is a point (or vector) in
   p-dimensional **feature space**
3. figure shows all possible pairwise projections on original variables



K-means

1. choose C (or k)
2. assign C class centres $\boldsymbol{\mu}_c$
3. calculate distance, e.g., $D^2_{Eic} = (\boldsymbol{x}_i - \boldsymbol{\mu}_c)^T(\boldsymbol{x}_i - \boldsymbol{\mu}_c)$ for all observations to all class centres, $i = 1, \ldots, N,\ c = 1, \ldots, C$
4. assign class c to $\boldsymbol{x}_i$ if distance smallest for class c
5. compute new class centres $\boldsymbol{\mu}_c$ (include only obs in class c)
6. iterate from third step

Initialization of $\boldsymbol{\mu}_c$

1. random observations within range of data
2. first C 'different enough' observations
3. based on PCA, e.g., uniformly distributed along first PC axis, or in plane spanned by two first PC axes
4. ...

Fuzzy c-means

1. choose C (or k)
2. assign C class centres $\boldsymbol{\mu}_c$
3. calculate distance, e.g., $D^2_{Eic} = (\boldsymbol{x}_i - \boldsymbol{\mu}_c)^T(\boldsymbol{x}_i - \boldsymbol{\mu}_c)$ for all observations to all class centres
4. assign degree of membership u_{ic} to $\boldsymbol{x}_i$ for all classes, e.g., $u_{ic} = (1/D^2_{Eic})/\sum_{j=1}^{C} 1/D^2_{Eij}$ leading to $\sum_{c=1}^{C} u_{ic} = 1$
5. compute new class centres (include all obs weighted by u_{ic}) $\boldsymbol{\mu}_c = \sum_{i=1}^{N} u_{ic}\boldsymbol{x}_i / \sum_{i=1}^{N} u_{ic}$
6. iterate from third step

Optimal number of clusters

- a good clustering has high between clusters variation (SS_B) and low within (among) clusters variation (SS_W)
- maximize the variance ratio criterion VRC

$$VRC(k) = \frac{SS_B(k)/(k-1)}{SS_W(k)/(N-k)}$$

sometimes called the Calinski-Harabasz clustering evaluation criterion

Probability

1. $0 \leq P(A_i) \leq 1$
2. $P(\Omega) = 1$
3. $P(\cup_i A_i) = \sum_i P(A_i),\; A_i \cap A_j = \emptyset,\; i \neq j$ (disjoint)
4. additivity
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
5. conditional probability

$$P(A \mid B) = P(A \cap B)/P(B),\; (P(B) > 0)$$

$$P(A \cap B) = P(A \mid B)\; P(B) = P(B \mid A)\; P(A)$$

Bayes' Rule

1. $A_1, \ldots, A_i, \ldots,$ disjoint and $\sum_i P(A_i) = 1$
2. $$P(B) = \sum_i P(B \cap A_i) = \sum_i P(B \mid A_i)\ P(A_i)$$
3. Bayes' rule
$$P(A_j \mid B) = P(A_j \cap B)/P(B) = \frac{P(B \mid A_j)\ P(A_j)}{\sum_i P(B \mid A_i)\ P(A_i)}$$
4. here
$$P(\omega_c|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|\omega_c)P(\omega_c)}{\sum_{j=1}^{C} P(\boldsymbol{X}|\omega_j)P(\omega_j)} \propto P(\boldsymbol{X}|\omega_c)\ P(\omega_c)$$
5. Bayes' classifier: choose maximum $P(\omega_c|\boldsymbol{X})$

MAP

1. $P(\omega_c|\boldsymbol{X}) \propto P(\boldsymbol{X}|\omega_c)\ P(\omega_c)$
2. $P(\omega_c)$ is prior (or a priori) probability
3. $P(\omega_c|\boldsymbol{X})$ is posterior (or a posteriori) probability
4. $P(\boldsymbol{X}|\omega_c)$ is the "likelihood", the data term, i.e., the conditional probability of the data given the class
5. max a posteriori probability: MAP estimation

Supervised: Gaussian $P(\boldsymbol{X}|\omega_c)$

1. Gaussian "likelihood" in 1-D
$$P(X|\omega_c) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma_c}\exp\left[-\frac{1}{2}\left(\frac{X-\mu_c}{\sigma_c}\right)^2\right]$$
σ_c^2 is variance for class c
2. Gaussian "likelihood" in p-D
$$P(\boldsymbol{X}|\omega_c) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}_c|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_c)\right]$$
$\boldsymbol{\Sigma}_c$ is $p \times p$ dispersion or covariance matrix for class c
$\boldsymbol{\Sigma}_c$ contains variances on diagonal and covariances off diagonal

Supervised: Gaussian $P(\boldsymbol{X}|\omega_c)$

1. MAP: max $P(\omega_c|\boldsymbol{X}) \Leftrightarrow$ max log-likelihood $\mathcal{L}_c(\boldsymbol{X})$ (use Bayes' rule)
$$\begin{aligned}\mathcal{L}_c(\boldsymbol{X}) &= \ln P(\omega_c|\boldsymbol{X}) \\ &= \ln P(\omega_c) + \ln P(\boldsymbol{X}|\omega_c) - \ln\sum_{j=1}^{C} P(\boldsymbol{X}|\omega_j)P(\omega_j)\end{aligned}$$
2. ln of sum in last term same for all c, drop it; insert Gaussian $\ln P(\boldsymbol{X}|\omega_c)$
$$\mathcal{L}_c(\boldsymbol{X}) \sim \ln P(\omega_c) - \frac{p}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_c| - \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_c)$$
3. drop $-\frac{p}{2}\ln 2\pi$
4. if equal priors: drop $\ln P(\omega_c)$

Supervised: Gaussian $P(\boldsymbol{X}|\omega_c)$

1. log-likelihood

$$\mathcal{L}_c(\boldsymbol{X}) \sim \ln P(\omega_c) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_c| - \frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_c)$$

   quadratic in $\boldsymbol{X}$: quadratic discriminant analysis
2. if equal dispersions, $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$: drop $-\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}$

$$\mathcal{L}_c(\boldsymbol{X}) \sim \ln P(\omega_c) + \boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \frac{1}{2}\boldsymbol{\mu}_c)$$

   linear in $\boldsymbol{X}$: linear discriminant analysis
3. min Mahalanobis distance?
4. min Euclidean distance?

Supervised: Confusion Matrix

1. or error matrix: measures quality of classification result
2. resubstitution or training/test

		Classified as										
		bare	urban	tailings	water	forest	tundra	poor veg	waste	wetland	Row total	Producer's accuracy (%)
Known class	bare	6628	1115	0	1	0	1	436	1857	563	10601	62.5
	urban	470	972	1	23	0	1	19	514	286	2286	42.5
	tailings	0	0	1076	0	0	0	0	0	0	1076	100.0
	water	8	17	0	4519	0	0	1	11	26	4582	98.6
	forest	0	0	0	0	1917	176	0	0	0	2093	91.6
	tundra	4	0	0	0	1973	22420	334	0	8	24739	90.6
	poor veg	180	40	0	1	0	91	4801	0	230	5343	89.9
	waste	30	27	0	0	0	0	0	865	29	951	91.0
	wetland	125	371	0	50	0	29	1136	129	6231	8071	77.2
Column total		7445	2542	1077	4594	3890	22718	6727	3376	7373	59742	
Consumer's accuracy (%)		89.0	38.2	99.9	98.4	49.3	98.7	71.4	25.6	84.5		

Comments

1. quadratic $\mathcal{L}$ and reject class
2. confusion matrix, learning/test samples, misclassification rate
3. histograms of all variables in all classes, derived features (e.g. $\sqrt{X}$, $\ln X$, products, ratios, principal components, local moments, ...)
4. calculate posterior $P(\omega_c|\boldsymbol{X})$ for each class
5. visualisation: Mahalanobis' distance $D^2_{Mic} = (\boldsymbol{X}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_c)$ as intensity or saturation, class as hue

Comments

- quadratic discriminant analysis: p elements in mean vector, $p(p+1)/2$ elements in dispersion matrix, problem for large p
- remedy
  - regularization
  - diagonal dispersion matrices for each class: p-dimensional Gaussian factors into product of p univariate Gaussians
  - linear discriminant analysis: all classes have same dispersion matrix
  - all classes have dispersion matrix equal to identity matrix
- Mahalanobis distance
- contour curves of constant Mahalanobis distance: for 2-D ellipse (broad, near circle vs thin, elongated), for p-D hyperellipsoid

Unsupervised: Gaussian mixture models, GMM

1. Bayes' rule: $P(\omega_c|\boldsymbol{x}_i) = K\ P(\boldsymbol{x}_i|\omega_c)P(\omega_c)$ with $1/K = \sum_{j=1}^{C} P(\boldsymbol{x}_i|\omega_j)P(\omega_j)$
2. **GMM**: Given some $u_{ic} = P(\omega_c|\boldsymbol{x}_i)$ with $\sum_{c=1}^{C} u_{ic} = 1$, calculate
3. $P(\omega_c) = \frac{1}{N}\sum_{i=1}^{N} u_{ic}$ (here the mixing proportion of class c)
   $\boldsymbol{\mu}_c = \frac{1}{NP(\omega_c)}\sum_{i=1}^{N} u_{ic}\boldsymbol{x}_i$
   $\boldsymbol{\Sigma}_c = \frac{1}{NP(\omega_c)}\sum_{i=1}^{N} u_{ic}(\boldsymbol{x}_i - \boldsymbol{\mu}_c)(\boldsymbol{x}_i - \boldsymbol{\mu}_c)^T$
4. $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ define $P(\boldsymbol{x}_i|\omega_c)$ which with $P(\omega_c)$ via Bayes' rule give a new $u_{ic} = P(\omega_c|\boldsymbol{x}_i)$ which in turn gives a new $P(\omega_c)$: iterate
5. example on Expectation Maximization (EM) algorithm
   E-step: calculate $P(\omega_c)$, $\boldsymbol{\mu}_c$, $\boldsymbol{\Sigma}_c$
   M-step: calculate $P(\omega_c|\boldsymbol{x}_i)$ in Bayes' rule

EM algorithm

- k-means and fuzzy c-means:

$$\max\ -\sum_{i=1}^{N}\sum_{c=1}^{C} u_{ic}^m(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{c}})^T(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{c}})$$

- GMM:

$$\max \sum_{i=1}^{N} \ln \sum_{c=1}^{C} P(\omega_c)P(\boldsymbol{x}_i|\omega_c)$$

(ln is optional)

Initialization of $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$

GMM initialization

1. - select observations at random as initial means
   - mixing proportions are uniform
   - initial covariance matrices are diagonal, elements on the diagonal are the variances
2. start with result from k-means or fuzzy c-means
3. ...

Hierarchical clustering

1. hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram
2. the tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level
3. this allows you to decide the level or scale of clustering that is most appropriate for your application
4. two extremes: every pixels is its own cluster vs entire image is one cluster

Other supervised methods

1. support vector machines, SVM
2. tree based methods, CART, random forests
3. artificial neural networks, ANN, CNN
4. ...

Classification take-away

- classification is the process of grouping observations (pixels or regions) into classes intended to represent different physical objects or types
- here, the production of a **thematic map** from (image) data with digital numbers representing for example reflected or emitted EM-radiation in different wavelength bands
- very many classification methods ranging from quite simple to highly advanced
- two major groups of methods: supervised and unsupervised
- use original data or derived (spatial) features; combine unsupervised and supervised

fin