Comp 775: Classification and Clustering

[Duda, Hart, Stork. *Pattern Classification*]

Marc Niethammer, Stephen Pizer
Department of Computer Science
University of North Carolina, Chapel Hill
Objectives for Statistics on Geometric Features

• **Classification**: e.g., diseased vs. normal, or for voxels: in tissue A vs. B vs. C
• Characterization of differences between classes
  • By **hypothesis testing** on null hypothesis: no difference
  • What kind of geometric differences?
  • Where are there geometric differences?
Many ways to define these decision boundaries.
Supervised/Unsupervised Classification

Supervised Classification

Requires training data from user for each class

Bayes Classifier
- Parametrized probabilities
- Parzen window
- Distance to data
- Parallelepiped
- Minimum Distance
- k-nearest neighbor

Threshold in separation direction
- Support vector machine
- Distance-weighted discrimination

Unsupervised Classification

Deriving the class assignments from the data

k-means
- Fuzzy c-means
- Hierarchical
Optimization with respect to feature

\[ \arg \max p(z | I) = \arg \max \log p(z | I) = \]

\[ z \quad z \]

\[ \arg \max [\log p(I | z) + \log p(z)]; \]

\[ z \quad \text{likelihood} \quad \text{prior} \]

Optimization with respect to class

\[ \arg \max P(c | z) = \arg \max \log P(c | z) = \]

\[ c \quad c \]

\[ \arg \max [\log p(z | c) + \log P(c)]; \]

\[ c \quad \text{likelihood} \quad \text{prior} \]
Posterior optimization via Bayes

• arg max \([\log p(I|z) + \log p(z)]\):
  \[z = \text{arg max } [\text{match of image } I \text{ to geometry } z + \text{ geometric typicality of } z]\]

• arg max \([\log p(z|c) + \log P(c)]\):
  \[c = \text{arg max } [\text{match of features } z \text{ to class } c + \text{ prevalence of } c]\]
Probability optimization criteria options for classification

• Posterior max (MAP): \( \arg \max_c p(\mathbf{z} | c) P(c) \)
• Likelihood max (ML): \( \arg \max_c p(\mathbf{z} | c) \)
  – \( = \) MAP with uniform prior
• Minimum risk
  – Requires costs of errors, benefits of successes
• Optimize ROC
  – Requires costs of errors, benefits of successes
• Minimum uncertainty (info. theoretic)
Techniques for estimation $p(z \mid c)$

- **Fit parametrized probability distribution**
  - E.g., Gaussian, with mean, principal directions, and principal variances as parameters
    - PCA

- **Unparametrized:** fit training histogram by smooth probability density function
  - Parzen windowing
Parzen Windowing

- Replace each training feature vector $\mathbf{z}$ by Gaussian($\mathbf{z}; 0, \sigma$) times $\mathbf{z}$

- Equivalently, convolve feature space histogram with Gaussian($\mathbf{z}; 0, \sigma$)
Parzen Window Classifier

Background histogram

Foreground histogram

Non-parametric estimation of the probability distributions. Can then be used for classification.
Multi-channel MR Image segmentation example

T1 weighted
Multi-channel MR Image segmentation example

T2 weighted
Multi-channel MR Image segmentation example

Segmented
Techniques for MAP classification of target case $\mathbf{z}$

- Compute $P(c|\mathbf{z}) \propto p(\mathbf{z}|c) \, P(c)$ for each class $c$ and choose class with largest
- Pre-compute decision regions and calculate decision region for $\mathbf{z}$
- Transform features $\mathbf{z}$ to $\mathbf{f}$ so that decision region boundaries are linear with normals $\mathbf{n}_i$ and thresholds $t_i$, and then judge using $\mathbf{f} \cdot \mathbf{n}_i < t_i$
Bayes Classifier

Probabilistic (typically parametric) modelling of the feature distribution. Pick the class with the highest posterior probability.

Image: N. Ray
Minimum Distance

Assignment to the class with the closest cluster center.

Image: L. van Vliet
K-nearest Neighbor (KNN)

Decision boundary governed by the labeling of the k-nearest neighbors.

Image: N. Ray
Parallelepiped

Classification by boxes.

Image: J. Chadwick
Separating Hyperplanes

- Normal to plane gives decision direction
- Plane gives threshold along normal for decision
  - Additional features as functions of other features to get nonlinear decision boundaries (kernel methods)
- Support Vector Machine (SVM) or Distance Weighted Discrimination (DWD)
Support Vector Machine

- Objective function to optimize over separating planes
  - Term for size of gap between classes
  - Term for misclassified cases
- Ultimately depends on the few cases nearest the gap plane ("support vectors")

Diagram by Bülent Üstün
Distance Weighted Discrimination (DWD) [Marron]

- Objective function to optimize
  - Term for distances of cases to hyperplane (sum of reciprocal squared distances)
  - Term for misclassified cases with distance weighting
- More robust than SVM when high dimension, low sample size
Unsupervised Training: Clustering

Finding clusters in feature space

- multi-spectral
- multi-channel
- multi-feature
- ...

Intra-cluster distances are minimized

Inter-cluster distances are maximized

Illustration: Tan, Steinbach, Kumar
k-means

Data

Illustration: Eytan Domany
k-means

Start with 3 cluster centers

Illustration: Eytan Domany
k-means

Assign points to closest cluster center

Illustration: Eytan Domany
k-means

Recompute cluster centers

Illustration: Eytan Domany
k-means

Repeat until convergence

Illustration: Eytan Domany
Fuzzy c-means

- Allows for class membership probability instead of hard class decisions
  - So in iteratively computing the cluster centers (means), you weight distances to the cluster centers, by the probability of a sample being in that class
  - Also in an iteration you are recomputing membership probabilities via relative distances to cluster centers

Image: Pham and Prince
Hierarchical, e.g., ISODATA

- Like $k$-means
- Allows for class merging or splitting, i.e., # of classes need not be pre-chosen