Pattern recognition
Advanced decision methods

Multiclass SVM

- Chapitre 5 -

Outline

Outline

Motivations

- Discrimination multihypothèse
- Many practical cases item theoretical difficulty: non-trivial extension of the results of learning theory to two classes
- Still an open topic

Multiclass Decision

General case:
A classification problem is defined by three elements:

- decision options that are hypothetically assumed all known
- the risk term to optimize
- performance constraints to satisfy.

Decision options:

$$\Psi = \{\psi_0, \psi_1, \ldots, \psi_{I-1}\}$$

$\mathcal{D}(\boldsymbol{x}) = k$ If $\boldsymbol{x}$ is assigned to the set of classes member of ψ_k.

Multiclass Decision

Risk term

$$r(\mathcal{D}) = \sum_{i=0}^{I-1} \sum_{j=0}^{nc-1} r_{ij} P_j P(\mathcal{D}(X) = i|\omega_j)$$

where r_{ij} the cost of assigning an observation of class ω_j to the subset of classes ψ_i.

Performance constrains:

Each constraint is defined by a cost function c_k with $k = 1..K$ and its associated bound γ_k:

$$\begin{aligned} c_k(\mathcal{D}) &\leq \gamma_k \\ \text{with: } c_k(\mathcal{D}) &= \sum_{i=0}^{I-1} \sum_{j=0}^{nc-1} c_{ij,k} P_j P(\mathcal{D}(X) = i|\omega_j) \end{aligned}$$

whire $c_{ij,k}$ is a real and $k = 1..K$.

Multiclass Decision

Problem

$$\begin{cases} \min_{\mathcal{D}} \; R(\mathcal{D}) \\ \text{with: } c_k(\mathcal{D}) \leq \gamma_k \quad \forall k = 1..K \end{cases}$$

Dual Problem

$$\max_{\boldsymbol{\mu} \geq 0} \left(\inf_{\mathcal{D}} Ł(\mathcal{D}, \boldsymbol{\mu}) \right)$$

$$\text{avec} \quad Ł(\mathcal{D}, \boldsymbol{\mu}) = R(\mathcal{D}) + \sum_{k=1}^{K} \mu_k \left(c_k(\mathcal{D}) - \gamma_k \right)$$

where $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_K]^T \in \mathbb{R}^{+K}$ is the vector of Lagrange multipliers; $\mathcal{D}$, the decision function which defines the partition $\mathcal{Z}$ de $\mathcal{X}$.

Optimal solution

Optimal rule

$\inf_{\mathcal{D}} Ł(\mathcal{D}, \boldsymbol{\mu})$ is given by the rule $\mathcal{D}_{\inf}$:

$$\mathcal{D}_{\inf}(\boldsymbol{x}, \boldsymbol{\mu}) = \underset{i,\ i=0..I-1}{\text{indicemin}}\ g_i(\boldsymbol{x}, \boldsymbol{\mu})$$

with:

$$g_i(\boldsymbol{x}, \boldsymbol{\mu}) = \sum_{j=0}^{nc-1} P_j P(\boldsymbol{x}|\omega_j)\left(r_{ij} + \sum_{k=1}^{K} \mu_k c_{ij,k}\right)$$

Concluding remarks

Identical rule

Bayes rule: solving the unconstrained problem of risk R' minimization:

$$R'(\mathcal{D}) = \sum_{i=0}^{I-1} \int_{\mathcal{D}(\boldsymbol{x})=i} P(\boldsymbol{x}) \sum_{j=0}^{nc-1} r'_{ij} P(\omega_j|\boldsymbol{x})\, d\boldsymbol{x}$$

with:

$$r'_{ij} = r_{ij} + \sum_{k=1}^{K} \mu_k^* c_{ij,k}$$

and $R'_i(\boldsymbol{x}) = g_i(\boldsymbol{x}, \boldsymbol{\mu}^*)$.

Bayes rule

Decision options:

$$\psi_k = H_k$$

for k from 1 to nc

Unconstrained rule

Bayes rule that is solution of the unconstrained minimization of risk R_B

$$R_B(\mathcal{D}) = \sum_{i=1}^{nc} \int_{\mathcal{D}(\boldsymbol{x})=i} \sum_{j=1}^{nc} r_{ij} P(\omega_j|\boldsymbol{x}) P(\boldsymbol{x}) d\boldsymbol{x}$$

with $\mathcal{D}_i$ the set of $\boldsymbol{x}$ such that $\mathcal{D}(\boldsymbol{x}) = i$ is defined by:

$$\mathcal{D}_i = \{x | \sum_{j=1}^{nc} r_{ij} P(\boldsymbol{x}|\omega_j) P_j\} \leq \sum_{j=1}^{nc} r_{kj} P(\boldsymbol{x}|\omega_j) P_j\} \forall k$$

Training - first solution

- Estimate $P\left(\boldsymbol{x}|\omega_j\right)$
- Estimate P_j
- Apply the Bayes rule by plugging the estimators

Principle

Not solve a more complex problem than the original problem, if the initial problem can be addressed directly.

Observation

Estimating densities is difficult (especially in large dimensional space).

Consequence

Try to define a partition.

Outline

Introduction

Principle:

- Decompose a multiclass problem into several 2 classes problems

Approaches:

- One against all
- One against one
- Error correcting codes
- Graphs

One against all

- A classifier per class: class k against the others
- The decision is made by selecting the rule for which $f(x)$ is the largest
- Problem: class size very unbalanced when learning ...
- Performance often good enough

One against one

- A classifier for each pair of classes
- We need C_K^2 classifiers for K classes
- Final decision based on majority vote (possibly weighted by $f(x)$)
- Outputs post-processing to estimate $P(\omega_i|f(x))$
  Platt proposes $\frac{1}{1+e^{\boldsymbol{A}f(x)+\boldsymbol{B}}}$

One against one - variation

- Basic classifier provide answers $\{-1, 0, 1\}$
- 0 for the other classes

Error correcting codes

Principle:

- Each class is characterized by a binary word of given size N
- Each 0, 1 is the result of a classification
- Each classification relates to a separation between groups of classes
- Each observation is classified N times
- Decision is taken according to the distances of the code word formed the codewords representing the classes

Remark:

- Effective if the bit errors are uncorrelated

Graphs and decision

Principle:

- Path between nodes
- Each node removes a class

Remarks:

- Learning as complex as 1 against 1
- sensitive to the order of nodes

Outline

M-SVM

Contexte:

- K classes
- A training set $\mathcal{A}_n = \{(x_i, y_i)\} \in (\mathcal{X} \times \{1 : K\})^n$

Goal:

- Find the hyperplane separators that minimize an objective function

$$J(f) = \lambda \|f\|^2_{\mathcal{H}} + \sum_{i=1}^{n} L(y_i, f(x_i))$$

$$\text{With} : \sum_{j=1}^{K} f_j = 0$$

- Consequence of representor theorem:

$$f_k(x) = \sum_{i=1}^{n} \alpha_{ik} K(x_i, x) + b_k$$

Weston and Watkins Model

Primal problem:

$$\min_f \left\{ \frac{1}{2} \sum_{j=1}^{K} \|w_j\|^2 + C \sum_{i=1}^{n} \sum_{j=1, j \neq y_i}^{K} \xi_{ij} \right\}$$

Avec $\quad \langle w_{y_i} - w_j, \phi(x_i) \rangle + b_{y_i} - b_j \geq 1 - \xi_{ij} \quad i = 1 : n, j = 1 : K, j \neq y_i$

$$\xi_{ij} \geq 0$$

The constraint $\sum_k w_k = 0$ is implicitly satisfied by the solution

Weston and Watkins Model

Dual Problem:

$$\min_\alpha \left\{\frac{1}{2}\alpha^T H\alpha - 1^T\alpha\right\}$$

$$\text{With} \qquad 0 \leq \alpha_{i,k} \leq C$$

$$\sum_{i|y_i=k} \sum_{l=1}^{K} \alpha_{il} - \sum_{i=1}^{m} \alpha_{ik} = 0$$

With $H = (h_{ik,jl})$ such that $h_{ik,jl} = K(x_i, x_j)(\delta_{y_i,y_j} - \delta_{y_i,l} - \delta_{y_j,k} + \delta_{k,l})$

Cramer and Singer Model

Primal problem:

$$\min_f \left\{ \frac{1}{2} \sum_{j=1}^{K} \|w_j\|^2 + C \sum_{i=1}^{m} \xi_i \right\}$$

$$\text{With} \quad \langle w_{y_i} - w_j, \phi(x_i) \rangle + \delta_{y_i,k} \geq 1 - \xi_i \quad i = 1 : m, j = 1 : K, j \neq y_i$$

$$\xi_i \geq 0$$

Training is focused on the most violated constraint for each sample. Leads to a more "compact" dual problem which enables a more effective implementation (by decomposition on the same principle as SMO)

Lee and al. Model

Both previous M-SVM does not converge to the Bayes classifier when the number of samples tends to ∞.
Lee and al. propose a universally convergent formulation.
Primal problem:

$$\min_f \left\{ \frac{1}{2} \sum_{j=1}^{K} \|w_j\|^2 + C \sum_{i=1}^{n} \sum_{j=1, j \neq y_i}^{K} \xi_{ij} \right\}$$

$$\text{With} \quad \langle w_j, \phi(x_i) \rangle + b_j \leq -\frac{1}{K-1} + \xi_{ij} \quad i = 1:n, j = 1:K, j \neq y_i$$

$$\xi_{ij} \geq 0$$

$$\sum_{j=1}^{K} w_j = 0, \sum_{j=1}^{K} b_j = 0$$

Consistency if cost converges in probability to Bayes risk
Strong consistency if convergence is ps (almost sure)
Universal Convergence: whatever P

On class SVM based solutions

Outline

Motivations