Kernel-based Methods for Unsupervised Learning

LEAR project-team, INRIA

Zaid Harchaoui

Lyon, Janvier 2011

Reminder : k nearest-neighbor

Nearest neighbor classifier

- Data : training set $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, +1\}$ for $i = 1, \ldots, n$
- Prediction : the predicted label y for a new $\mathbf{x}$ is the most popular (majority vote) label of the nearest neighbors of $\mathbf{x}$ in the training set

Strength and weakness

- Strength : "plug-and-play" and wide applicability to a large class of learning problems (even those difficult to cast into the supervised/unsupervised taxonomy)
- Weakness : no explicit regularization penalty allowing to control the complexity of the classifier (especially for high-dimensional input space)

Teaser on supervised learning

The need for regularization

- Data : (training) set $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$
- Minimize w.r.t. $f \in \mathcal{H}$

$$\|f\|^2 \quad + \frac{C}{n} \sum_{i=1}^{n} \ell(y_i, f(\mathbf{x}_i))$$

Regularization penalty data-fitting

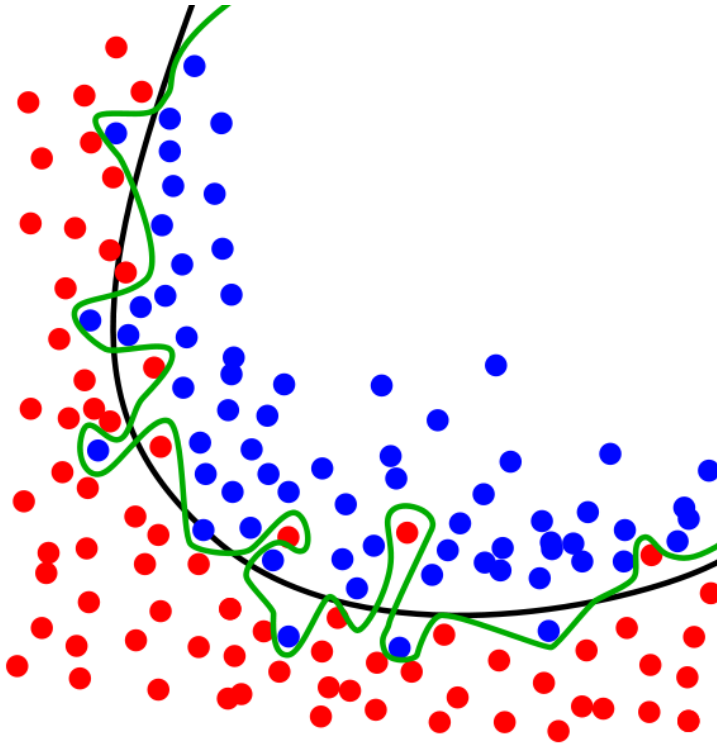
Extreme regimes

- $C \to \infty$: learning by heart $\to$ *overfitting*
- $C \to 0$: no learning at all $\leftarrow$ *underfitting*
- in practice : C is set so as to minimize average error on several random splits of the dataset into training and validation sets $\to$ cross-validation

The need for regularization

the overfitting phenomenon

Data : $(\mathbf{x}_i, y_i)_{i=1,\ldots,n} \in \mathbb{R}^2 \times \{-1, +1\}$



Supervised classification

Problem

- Data : $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, +1\}$ for $i = 1, \ldots, n$
- Goal : learn a binary-valued function $g(\cdot)$ such that g yields low error when it comes to predict y for a new unseen $\mathbf{x}$

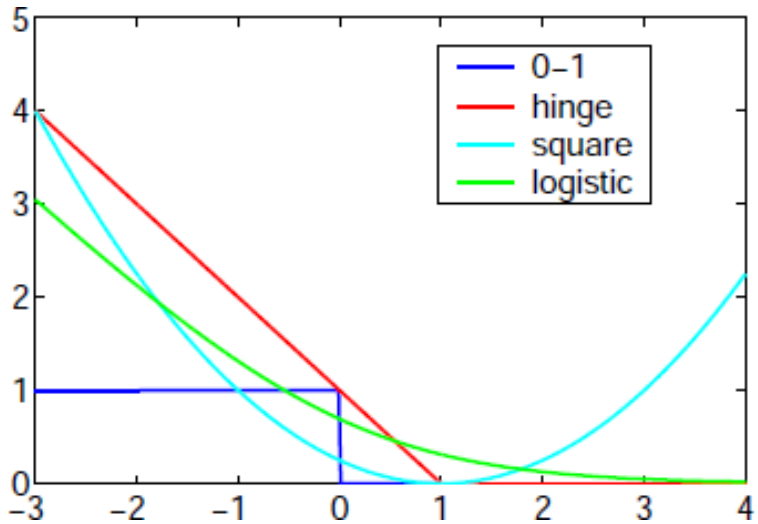
Relaxation

- Relax to continuous : too hard to learn a binary-valued function $\rightarrow$ learn a real-valued function f and take its sign to predict y
- Relax to convex : too hard to minimize a non-convex non-smooth objective couting misclassification errors $\rightarrow$ minimize a convex upper bound on the original objective

Convex loss function for supervised classification

the overfitting phenomenon

- Data : $(\mathbf{x}_i, y_i)_{i=1,\dots,n} \in \mathcal{X} \times \{-1, +1\}$
- "True loss" : misclassification loss $\mathbf{1}\{yf(\mathbf{x}) < 0\}$
- Convex surrogate loss : can be written in the form $\ell(y, f) = \ell(yf)$



Support Vector Machine

Characterization

- Function : $f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x})$
- Regularization penalty : RKHS norm $\|f\|^2 = \alpha^T \mathbf{K} \alpha$
- Loss function : *hinge loss* $\ell(y, f) = \max(0, 1 - yf)$, convex but non-smooth $\rightarrow$ sparsity of α
- Support Vectors : the datapoints $\mathbf{x}_i$ to which correspond nonzero α_i-s.