Kernel-based methods

26 Mar 2023

Table of contents

Kernel-based methods

- depend only on some *distance function* induced between pairs of points
- kernel function k
  - classification: $\hat{y}_i(\mathbf{x}') = textrmsign \sum w_i y_i k(\mathbf{x}_i, \mathbf{x}_i')$
  - regression: (the same but without the sign()!)
- **low-rank approximations** of $k()$

Kernel smoothers

- kernel density estimation
- Nadaraya-Watson kernel regession

Separating hyperplanes

- ESL section 4.5
- Regress $\mathbf{y} \in \{-1, 1\}$ on $\mathbf{x}$: solve for $\mathbf{X}\beta = 0$ (write as $\beta_0 + \beta^\top \mathbf{x} = 0$, i.e. separate intercept)
- (equivalent to linear discriminant analysis)
- Rosenblatt's algorithm
  - $(\mathbf{X}\beta)/||\beta||$ is the signed distance to the separating plane
  - minimize $-\sum i \in M y_i(\mathbf{X}\beta)$ (sum of misclassified distances)
  - gradient $= -\sum(y_i x_i)$
  - stochastic gradient descent* (pointwise): adjust β by $\rho \mathbf{y}_i X_i$ at each step
- elegant but not practical (non-unique, slow, non-convergent if not separable)
- → penalized version in a larger basis space
- $\text{argmin}(\beta)\frac{1}{2}||\beta||^2$ subject to $y_i(\mathbf{X}\beta) \geq 1$
- "standard" convex optimization problem

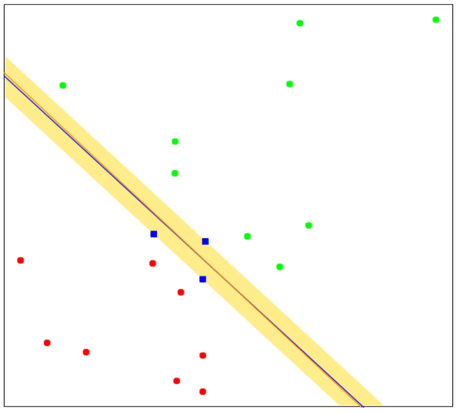

FIGURE 4.16. *The same data as in Figure 4.14. The shaded region delineates the maximum margin separating the two classes. There are three support points indicated, which lie on the boundary of the margin, and the optimal separating hyperplane (blue line) bisects the slab. Included in the figure is the boundary found using logistic regression (red line), which is very close to the optimal separating hyperplane (see Section 12.3.3).*

support vector machines for the non-separable case

- ESL chapter 12
- $y_i(X_i\beta \geq M(1-\xi_i)$
- linear loss function on misclassification distances + L2 penalty
- or $\min \frac{1}{2}||\beta||^2 + C\sum \xi_i$
- C is the hyperparameter
- quadratic programming problem

SVMs and kernels (ESL 12.3)

- alternative formulation

$$\begin{aligned} f(x_i) &= X_i^\top \beta + \beta_0 \\ &= \sum \alpha_j y_j \left\langle h(x_i), h(x_j) \right\rangle + \beta_0 \end{aligned}$$

where α_i is a different parameterization * $\langle h(.), h(.) \rangle$ is a **kernel function** * linear SVM finds a separating hyperplane based on distances * polynomial distance: $(1 + \left\langle x_i, x_j \right\rangle)^d$ * polynomial d for n inputs (plus intercept) gives rise to a $C(n+2, d)$-dimensional space * **radial basis function** $\exp\left(-\gamma||x_i - x_j||^2\right)$ * infinite-dimensional (think of Taylor expansion) * **length scale** $1/\gamma$

SVMs for regression

- fits a loss function $\max(0, |r| - \epsilon)$

kernels

- "kernel trick" works very generally, but only for L2 penalty
- ESL 12.3.7: cost of optimizing via kernel is $O(N^2)$ not $O(MN^2)$ (where N is number of training points, M is dimension of the feature space)

Gaussian processes

- Rasmussen and Williams (2005)
- motivated by Bayesian context, or from classical **geostatistics** (kriging)
- interpolation vs. approximation
- "Under the assumption of Gaussian observation noise the computations needed to make predictions are tractable and are dominated by the inversion of a n × n matrix."
- zero-mean Gaussian prior: $\mathbf{w} \sim N(0, \Sigma_p)$

Rasmussen, Carl Edward, and Christopher K. I. Williams. 2005. *Gaussian Processes for Machine Learning.* Cambridge, Mass: The MIT Press.