Machine Learning
Exercise 10

Marc Toussaint
Machine Learning & Robotics lab, U Stuttgart
Universitätsstraße 38, 70569 Stuttgart, Germany

June 27, 2016

1 Gradient Boosting for classification

Consider the following *weak learner* for classification: Given a data set $D = \{(x_i, y_i)\}_{i=1}^n, y_i \in \{-1, +1\}$, the weak learner picks a single i^* and defines the discriminative function

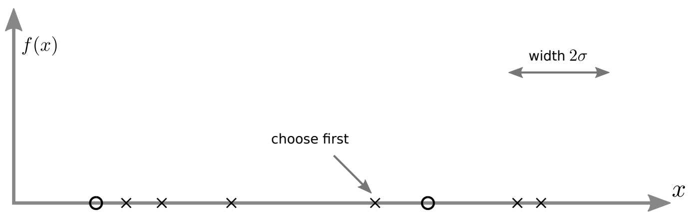
$$f(x) = \alpha e^{-(x-x_{i^*})^2/2\sigma^2} ,$$

with fixed width σ and variable parameter α. Therefore, this weak learner is parameterized only by i^* and $\alpha \in \mathbb{R}$, which are chosen to minimize the neg-log-likelihood

$$L^{\text{nll}}(f) = -\sum_{i=1}^n \log \sigma(y_i f(x_i)) .$$

a) Write down explicitly a pseudo code for gradient boosting with this weak learner. By "pseudo code" I mean explicit equations for every step that can directly be implemented in Matlab. This needs to be specific for this particular learner and loss.

b) Here is a 1D data set, where $\circ$ are 0-class, and $\times$ 1-class data points. "Simulate" the algorithm graphically on paper.



c) If we would replace the neg-log-likelihood by a hinge loss, what would be the relation to SVMs?

2 Sum of 3 dices

You have 3 dices (potentially fake dices where each one has a different probability table over the 6 values). You're given all three probability tables $P(D_1)$, $P(D_2)$, and $P(D_3)$. Write down a) the equations and b) an algorithm (in pseudo code) that computes the conditional probability $P(S|D_1)$ of the sum of all three dices conditioned on the value of the first dice.

3 Product of Gaussians

A Gaussian distribution over $x \in \mathbb{R}^n$ with mean μ and covariance matrix Σ is defined as

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{|\,2\pi\Sigma\,|^{1/2}} \, e^{-\frac{1}{2}(x-\mu)^\top \, \Sigma^{-1} \, (x-\mu)}$$

Multiplying probability distributions is a fundamental operation, and multiplying two Gaussians is needed in many models.

a) Prove

$$\mathcal{N}(x \,|\, a, A)\; \mathcal{N}(x \,|\, b, B) \;=\; c\; \mathcal{N}(x \,|\, (A^{\text{-}1} + B^{\text{-}1})^{\text{-}1}(A^{\text{-}1}a + B^{\text{-}1}b), (A^{\text{-}1} + B^{\text{-}1})^{\text{-}1})\ .$$

where c are terms independent of x.

b) Now also compute the normalization c as a function of a, A, b, B.

Note: The so-called canonical form of a Gaussian is defined as $\mathcal{N}[x \,|\, \bar a, \bar A] = \mathcal{N}(x \,|\, \bar A^{\text{-}1}\bar a, \bar A^{\text{-}1})$; in this convention the product reads much nicher: $\mathcal{N}[x \,|\, \bar a, \bar A]\; \mathcal{N}[x \,|\, \bar b, \bar B] \propto \mathcal{N}[x \,|\, \bar a + \bar b, \bar A + \bar B]$. You can first prove this before proving the above, if you like.