Machine Learning, Fall 2014, Exercises 3

Please return your solutions on Friday Oct 9 in the lecture. Legible handwriting or typeset printouts highly appreciated.

This time, paper-and-pencil tasks! If you wish you may team up in groups of 2 (not more) and hand in a single solution sheet per team.

Problem 1. (in sum, 70 points). A European-wide operating insurance company wants to have a predictor for the creditworthiness of its customers. The database of the company's $N = 200{,}000$ past loan contracts has for each contract a 60-dimensional feature vector $\mathbf{x}_i$, with a known binary class label $d_i \in \{1, 2\}$, where class 1 is the class of creditworthy customers (paid back loan) and class 2 is the class of non-creditworthy customers. Some of the 60 entries in $\mathbf{x}_i$ are symbolic (e.g., "male", "female"), others are numerical (e.g., "last year's income").

a. (35 points) Give a probabilistic interpretation of the database $(\mathbf{x}_i, d_i)_{i = 1...N}$. Deliverable: Describe in words a possible probability space Ω (what are the $\omega \in \Omega$?), the random variable(s), and the observation space E (what are its elements?). Take into account that a European citizen might have several loan contracts with the company.

b. (35 points) Obviously, with 60-dim input data, deriving a classificator runs into the curse of dimensionality. One (rather brutal) way to fight this curse is to choose from the 30 customer features a small subset, which is exclusively used for the prediction, and just ignore all others. Describe in words a greedy search scheme to select from the 30 features a small subset which yields good prediction performance. You may assume that the company's statistician has, for each (small) subset $\mathbf{x}'$ a method to train a predictor $f_{\mathbf{x}'}$: $\mathbf{x}' \mapsto \hat{d}$.

Problem 2. (30 points). Consider a binary, one-dimensional classification task, where observations from class C_1 are distributed by $N(0,1)$ and from class C_2 by $N(1,1)$ (where $N(\mu, \sigma^2)$ is the normal distribution with mean μ and variance σ^2). Let $P(C_1) = 0.25$. At which point lies the optimal decision boundary b? Give a formula and a rough numerical estimate (use $\ln 3 \approx 1.1$), and a rough plot of $p(x \mid C_1)\, P(C_1)$ and $p(x \mid C_2)\, P(C_2)$. Remember that the pdf of a one-dimensional Gaussian is given by

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$