T-61.5040 Oppivat mallit ja menetelmät
T-61.5040 Learning Models and Methods
Pajunen, Viitaniemi

Solutions to exercise 2, 26.1.2007

Problem 1.

Localized basis functions are one method of nonparametric modeling.

i)

$$E\{y|x\} = \int yp(y|x)\,dy$$

Use

$$p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x,y)}{\int p(x,y)\,dy}$$

to get the regression function

$$E\{y|x\} = \frac{\int yp(x,y)dy}{\int p(x,y)dy}.$$

ii) Place $1/N$-th of the probability mass at each observation, that is, $K(x-x_i, y-y_i)/N$ at (x_i, y_i). Then the joint density of the observations is

$$p(x,y) = 1/N \sum_i K(x-x_i, y-y_i)$$

which is indeed a density function because it integrates to 1 and it is nonnegative.

iii) Now insert the formula of $p(x,y)$ into the regression function $E\{y|x\}$ obtained in part i):

$$\begin{aligned} E\{y|x\} &= \frac{\int yp(x,y)dy}{\int p(x,y)dy} \\ &= \frac{\int y \sum_i K(x-x_i, y-y_i)dy}{\int \sum_i K(x-x_i, y-y_i)dy} \\ &= \frac{\sum_i \left(K_x(x-x_i)\int yK_y(y-y_i)dy\right)}{\sum_i \left(K_x(x-x_i)\int K_y(y-y_i)dy\right)} \\ &= \frac{\sum_i K_x(x-x_i)y_i}{\sum_i K_x(x-x_i)} \end{aligned}$$

In the last equation we used the fact that $K_y(y-y_i)$ is the density function of the Gaussian distribution $N(y_i, 1)$, and formula $\int yK_y(y-y_i)dy$ gives the expectation y_i of this distribution.

Note that using the regression function $E\{y|x\}$ as an estimate for the value of y given x is a justified choice if the least mean squared error is considered.

Problem 2.

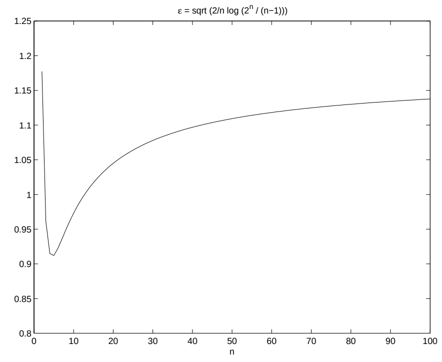
i) The classifiers are defined by selecting the index $i \in \{1, 2, \ldots, n-1\}$ and deciding whether x_i is in class 0 or 1. This gives a total of $2n-2$ different "nice" classifiers.

ii) The assumptions made above mean that the fraction of "nice" classification problems is $(2n-2)/2^n$. In Problem 5 / Exercise 1, we computed an upper bound for the fraction of problems where the performance of any two methods differ more than ϵ. Here we use the given upper bound $2e^{-\epsilon^2(n/2)}$ (*).

Consider A=guessing and B=potentially very good classifier. Suppose you want to claim that B is much better than A, meaning that ϵ is large. Then you use (*) to see how small the fraction of problems must be for this to be true. Increasing ϵ, the upper bound gets smaller. At some ϵ, the upper bound is equal to $(2n-2)/2^n$. If this ϵ is small, you can conclude that in the set of "nice" problems, B is not much better than guessing. If ϵ is large, then you can conclude that B may be much better than guessing. Compute the ϵ where the upper bound equals $(2n-2)/2^n$:

$$\begin{aligned} \frac{2n-2}{2^n} &= 2e^{-\epsilon^2(n/2)} \\ \epsilon &= \sqrt{\frac{2}{n}\log(\frac{2^n}{n-1})} \end{aligned}$$

This ϵ is typically larger than 1, as seen in the figure. Therefore the set of "nice" problems is small enough that (*) cannot limit the performance of B.



Comments: The assumption made above (the problem is "nice") is often reasonable in classification problems. However, for regression problems it would seem to be quite strong. Yet both regression and classification are exactly the same problems: there are inputs x and corresponding outputs y. In classification, loose assumptions such as made above seem to imply much more on the possibility of having useful learning methods.

Note that all this requires that all errors are equally bad: in regression one often prefers small errors to large errors, but in classification all errors are often equal. We avoided these questions by choosing the outputs to be binary.

Problem 3.

i) There are two points and therefore the line can be exactly fitted. The equations are

$$y_1 = \hat{\mu} + \hat{\beta}, \quad y_0 = \hat{\beta} \quad \implies \hat{\beta} = y_0 \text{ and } \hat{\mu} = y_1 - y_0$$

The prediction at $x = 2$ is $\hat{y}_2 = 2\hat{\mu} + \hat{\beta} = 2y_1 - y_0$. The distribution of the predicted value is Normal. The mean value is $2E[y_1] - E[y_0] = 2\mu + \beta$, and the variance is $Var(2y_1 - y_0) = 4Var(y_1) + Var(y_0) = 5$ (here we have taken the noise term n to be generated independently for each observation). So $\hat{y}_2 \sim N(2\mu + \beta, 5)$. Since the true value at $x = 2$ is $2\mu + \beta$, the mean-square error is 5.

ii) The constant minimizing the mean-square error is the average $y_0/2 + y_1/2$. This is at the same time the prediction at all inputs x. It is Normally distributed with mean $1/2E[y_0] + 1/2E[y_1] = 1/2\beta + 1/2(\mu + \beta) = 1/2\mu + \beta$ and variance is $1/4Var(y_0) + 1/4Var(y_1) = 1/2$. Therefore the prediction $\hat{y}_2$ has a distribution $N(1/2\mu + \beta, 1/2)$. The mean squared error is $E([\hat{y}_2 - 2\mu - \beta]^2) = E[z^2]$ where we denote $z = \hat{y}_2 - 2\mu - \beta$. z has normal distribution $N(-3/2\mu, 1/2)$ and therefore it is easy to calculate $E[z^2] = Var(z) + E[z]^2 = 1/2 + (-3/2\mu)^2$.

If this mean-square error is less than 5, then it may make sense to use a constant regression function *even if you know that the true model is linear.* If $\mu = 1$, then the MSE for the constant model is $11/4 < 5$. Overfitting can happen when there is not enough data.

Problem 4.

i)

$$\begin{aligned} P(|a_1 - b_1| \le x) &= 1 - P(|a_1 - b_1| > x) \\ &= 1 - 2P(a_1 - b_1 > x) \\ &= 1 - 2\int_0^{1-x} \left(\int_{b_1+x}^1 1 da_1 \right) db_1 \\ &= 1 - 2\int_0^{1-x} (1 - b_1 - x) db_1 \\ &= 1 - 2|_0^{1-x} \left(b_1 - \frac{1}{2}b_1^2 - xb_1 \right) \\ &= 1 - 2\left(1 - x - \frac{1}{2}(1-x)^2 - x(1-x) \right) \\ &= 1 - 2\left(\frac{1}{2}x^2 - x + \frac{1}{2} \right) \\ &= 1 - x^2 + 2x - 1 \\ &= x(2 - x). \end{aligned}$$

ii) z is the maximum of variables $|a_i - b_i|$. Therefore

$$P(z \le x) = \prod_{i=1}^d P(|a_i - b_i| \le x) = x^d(2 - x)^d.$$

iii) As a minimum of $n - 1$ variables z_j, we have

$$\begin{aligned} P(w \le x) &= 1 - \prod_{j=2}^n (1 - P(z_j \le x)) \\ &= 1 - (1 - x^d(2 - x)^d)^{n-1} \\ E(w) &= \int_0^1 1 - P(w \le x)dx = \int_0^1 \left(1 - x^d(2 - x)^d\right)^{n-1} dx. \end{aligned}$$

iv)

$$\begin{aligned} E(w) &= \int_0^1 (1 - 2x + x^2)^{n-1} dx \\ &= \int_0^1 (x - 1)^{2n-2} dx \\ &= \int_{-1}^0 x^{2n-2} dx \\ &= \frac{1}{2n - 1}. \end{aligned}$$

v) We are evaluating the integral

$$E(w) = \int_0^1 \left(1 - x^d(2 - x)^d\right)^{n-1} dx.$$

Monotone convergence lemma says that $\lim_{d\to\infty} \int f_d(x)dx = \int \lim_{d\to\infty} f_d(x)dx$ when $f_d(x) \ge 0$ and $f_{d+1}(x) \ge f_d(x)$ for all d and $x \in [0, 1]$ and $\lim_{d\to\infty} f_d(x)$ exists. We take the integrand $(1 - x^d(2 - x)^d)^{n-1}$ to be our function f. Since $0 \le x(2 - x) < 1$ for $0 \le x < 1$,

$$\lim_{d\to\infty} \left(1 - x^d(2 - x)^d\right)^{n-1} = 1$$

when $0 \le x < 1$ and zero for $x = 1$. Furthermore, for fixed x in the interval f is non-negative and does not decrease as d grows.

The conditions of the lemma are thus fulfilled and we get

$$E(w) = \lim_{d\to\infty} \int_0^1 \left(1 - x^d(2 - x)^d\right)^{n-1} dx E(w) = \int_0^1 \lim_{d\to\infty} \left(1 - x^d(2 - x)^d\right)^{n-1} dx = \int_0^1 1dx = 1.$$

Note: the calculation was based on the limit of the integrand being one for a fixed n. The same can be shown to hold even if the number of points n is any polynomial function of the dimension d (but the corresponding limiting procedure is a bit more complicated). We can thus say that a polynomial number of points is not enough to densely cover the cube as the dimension of the space increases.

vi) $E(w) \approx (\frac{1}{n})^{1/d}$ since this gives the volume $\frac{1}{n}$ for one small cube, and there are n of them.

Also this approximation tends to unity if n is any polynomial of d.