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

Exercises 2, 26.1.2007

Problem 1.

You are given a set of observations (x_i, y_i), $i = 1, \ldots, N$. If you want to predict y using observed input x, the predicted value $\hat{y}$ that minimizes the MSE $\mathrm{E}((\hat{y}-y)^2)$ is the *conditional expectation* $\mathrm{E}(y|x)$.

i) To demonstrate that knowing the joint distribution $p(x, y)$ allows us to solve a regression problem, compute $\mathrm{E}(y|x)$ as a function of $p(x, y)$. Recall that $p(y, x) = p(y|x)p(x)$ and $p(x) = \int p(x, y)dy$.

ii) Estimate the joint distribution $p(x, y)$ as a sum of localized density functions

$$K(x - x_i, y - y_i) = K_x(x - x_i)K_y(y - y_i)$$

where x_i, y_i are constants and K_x, K_y are also density functions:

$$\begin{aligned} K(x, y) &= (2\pi)^{-1} \exp(-0.5(x^2 + y^2)) \\ K_x(x) &= (2\pi)^{-1/2} \exp(-0.5x^2) \\ K_y(y) &= (2\pi)^{-1/2} \exp(-0.5y^2). \end{aligned}$$

iii) Now estimate $\mathrm{E}(y|x)$ using the above kernel estimate of the density function $p(x, y)$. Can you interpret the result geometrically?

Problem 2.

Consider a set of distinct input points $x_1, \ldots, x_n$ and all possible outputs $y_i \in \{0, 1\}$. This results in 2^n different regression functions. Assume that the points x_i are ordered, i.e. $x_i < x_j$ when $i < j$.

i) Calculate the number of different classifiers if we assume that all classification problems are "nice": both classes 0 and 1 are clustered and there is exactly one $i \in \{1, 2, \ldots, n - 1\}$ for which x_i and x_{i+1} belong to different classes.

ii) In Problem 5 / Exercise 1 we studied the probability for the difference between two classifiers. Now let us use the upper bound for difference in fraction of errors: $P(\text{difference} \geq \epsilon) < 2e^{-\epsilon^2(n/2)}$ where n is the number of input points. Compute the fraction of "nice" problems from i), and use the upper bound to see how good a classifier can possibly be on "nice" problems.

Problem 3.

Consider two observations, y_0 at an input $x = 0$, and y_1 at an input $x = 1$. Assume that you know the correct model and it is linear: $y = \mu x + \beta + n$, where n has a Normal distribution $N(0, 1)$ (1 is the noise variance).

Assume that you fit a line to these observations by minimizing the mean squared error, and then use the linear model to predict the output at $x = 2$.

i) What is the mean squared error of the prediction at $x = 2$?

ii) Repeat part i), but instead of a line, fit a *constant regression function.* What can you conclude if $\mu = 1$?

Problem 4.(demo)

In high dimensions, observations tend to be far away from each other. To cover a high-dimensional space, lots of data are needed.

Consider a d-dimensional unit hypercube $[0, 1]^d$. Assume it contains n points which have been randomly drawn from the uniform distribution. We will consider the L_∞ norm

$$\|a - b\|_\infty = \max\{|a_1 - b_1|, \ldots, |a_d - b_d|\}$$

where $a = (a_1, \ldots, a_d)$ and $b = (b_1, \ldots, b_d)$, both uniformly distributed.

i) What is the probability $P(|a_1 - b_1| \leq x)$, i.e. what is the cumulative distribution function of $|a_1 - b_1|$?

ii) What is the probability $P(z \leq x)$ where $z = \|a - b\|_\infty$?

iii) Denote by z_j the $L_\infty - distance$ from point 1 to point j where $j = 2, \ldots, n$. Denote by w the distance from point 1 to the closest point, i.e. $w = \min_j z_j$. What is $P(w \leq x)$? Write $\mathrm{E}(w)$ as an integral over x.

iv) Solve the expected distance $\mathrm{E}(w)$ when $d = 1$.

v) What is $lim_{d \to \infty} \mathrm{E}(w)$ when number of points n is fixed?

vi) Roughly approximate $\mathrm{E}(w)$ by filling the hypercube with n identical small cubes and computing the side length of the small cube.

Lots of hints:

- The maximum z of a finite set $(w_1, w_2, \ldots, w_n)$ of random variables has the distribution
$$P(z \leq x) = \prod_i P(w_i \leq x).$$
- The minimum z of a finite set $(w_1, w_2, \ldots, w_n)$ of random variables has the distri-

bution

$$P(z \leq x) = 1 - \prod_i (1 - P(w_i \leq x)).$$

- To compute expectation of a non-negative random variable using the distribution function, use

$$E(w) = \int_0^\infty 1 - P(w \leq x)dx.$$

- Monotone convergence lemma:

$$\lim_{d \to \infty} \int f_d(x)dx = \int \lim_{d \to \infty} f_d(x)dx$$

when $f_d(x) \geq 0$, $f_{d+1}(x) \geq f_d(x)$ and $\lim_{d \to \infty} f_d(x)$ exists for all d and x.