Ludwig-Maximilians-Universitaet Muenchen
Institute for Informatics
Prof. Dr. Volker Tresp
Gregor Jossé
Johannes Niedermayer

28.05.2015

Machine Learning and Data Mining

Summer 2015

Exercise Sheet 6

Presentation of Solutions to the Exercise Sheet on the 03.06.2015

Exercise 6-1 Maximum Likelihood Estimator

Suppose when soccer players train penalty kicks, each training session ends after their first miss, since they are demotivated. Let the probability of a miss be $p \in [0{,}1]$. Then, the probability for exactly x_i hits ($i \in \{1, \ldots, N\}$) before the first miss can be modeled using the geometric distribution:

$$P(x_i) = p \cdot (1-p)^{x_i}$$

(a) Following a frequentist approach, determine the maximum likelihood estimator p^{ML} for an i.i.d. (independent identically distributed) population of N soccer players, which hit $x_i \in \{0, 1, \ldots, \infty\}$ times before their first miss.

Possible Solution:

Log-Likelihood-function:

$$L(p) = \prod_{i=1}^{N} p(1-p)^{x_i} \qquad \text{(i.i.d.)}$$

$$l(p) = \sum_{i=1}^{N} \log\left(p(1-p)^{x_i}\right) = \sum_{i=1}^{N} \log p + \sum_{i=1}^{N} \log(1-p)^{x_i} = N \log p + \sum_{i=1}^{N} x_i \cdot \log(1-p) =$$

$$= N \log p + (\log(1-p)) \cdot \left(\sum_{i=1}^{N} x_i\right)$$

Derive and set $= 0$:

$$\frac{\partial l(p)}{\partial p} = \frac{N}{p} + \frac{1}{1-p} \cdot (-1) \cdot \sum_{i=1}^{N} x_i = \frac{N}{p} - \frac{1}{1-p} \cdot \sum_{i=1}^{N} x_i \overset{!}{=} 0$$

$$\frac{N}{p} = \frac{1}{1-p} \cdot \sum_{i=1}^{N} x_i$$

$$\frac{1-p}{p} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i = \frac{1}{p} - 1$$

$$\Rightarrow p^{\text{ML}} = \frac{1}{1 + \frac{1}{N} \cdot \sum_{i=1}^{N} x_i}$$

(b) Consider the following dataset $\mathbf{X} = \begin{array}{|c|c|}\hline 7 & 2 \\ \hline\end{array}$ for $N = 2$ training sessions. Compute the probability of a miss.

Possible Solution:

Simply: $\frac{\text{\#sessions}}{\text{\#tries}} = \frac{N}{N+\sum_i x_i} = \frac{2}{2+9} = \frac{2}{11}$.
Or with the ML estimator from a): $p^{\text{ML}} = \frac{1}{1+\frac{1}{N}\sum_{i=1}^{N} x_i} = \frac{1}{1+0{,}5\cdot 9} = \frac{1}{5{,}5} = \frac{2}{11}$.

Exercise 6-2 Frequentist versus Bayesian Statistics

Consider this – rather pathological – example to illustrate the difference between frequentist and bayesian statistics: Alice and Bob play a game in which the first person to get 6 points wins. The points are scored in the following way: A referee is standing at a pool table Alice and Bob cannot see. Before the game begins, the referee rolls a ball onto the table coming to rest at a random position. Each point scored is decided by the referee rolling another ball. If the ball comes to rest left of (the middle of) the initial ball, Alice scores, if it comes to rest right, Bob scores. The players know nothing but who scored a point. If the portion left of the initial ball is denoted as p, it is obvious that the probability of Alice scoring a point is p.

Now, consider the following situation within the game: Alice has 5 points, Bob has 3. Let us investigate the probability of Bob winning.

(a) Assume that the initial ball came to rest such that $p = 2/3$. What is the probability that Bob wins?

Possible Solution:

Obviously, $P(\text{Bob wins}) = (1-p)^3 = 1/27$

(b) Unfortunately, we do not know p – we only have some data we can try to estimate it from. Follow a *frequentist* approach: compute the maximum likelihood estimator for p and the probability of Bob winning.

Possible Solution:

$$l(p) = \log L(p) = \sum_{i=1}^{8} \log P_p(x_i) = 3\log(1-p) + 5\log(p)$$

$$\begin{aligned}\frac{\partial l}{\partial p} &= \frac{-3}{1-p} + \frac{5}{p} \overset{!}{=} 0\\ &\Leftrightarrow 5 - 5p = 3p\\ &\Leftrightarrow p^{\text{ML}} = 5/8\end{aligned}$$

Therefore, $P(\text{Bob wins}) = (1-p^{\text{ML}})^3 \approx 0.053$.

(c) Now, let us follow a *bayesian* approach: We know that p is only dependent on the position of the initial ball which is uniformly distributed on the table, i.e., $unif[0{,}1]$. Note that we compute the expected probability of Bob winning, as p itself is now drawn from a distribution. *Hint:* You will need the beta function:

$$B(x{,}y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{(x-1)!(y-1)!}{(x+y-1)!}$$

Optionally, you may implement the experiment and empirically compute the solution.

Possible Solution:

The expected probability of Bob winning is an integral over all possible values of p:

$$E(\text{Bob wins}) = \int_0^1 (1-p)^3 \cdot P(p \mid A=5,B=3)dp$$

In order to compute the posterior probability $P(p \mid A=5,B=3)$ we need Bayes' formula:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)} = \frac{P(Y \mid X)P(X)}{\sum_Z P(Y \mid Z)P(Z)}$$

where the Z are a partition of the event space. In this particular case:

$$P(p \mid A=5,B=3) = \frac{P(A=5,B=3 \mid p)P(p)}{\int_0^1 P(A=5,B=3 \mid q)P(q)dq}$$

We can compute $P(A=5,B=3 \mid p)$:

$$P(A=5,B=3 \mid p) = \frac{8!}{5!3!}p^5(1-p)^3 = 56p^5(1-p)^3 \quad (\star)$$

And $P(q)$ is a independent of q, we can pull it out of the integral and cancel it out. It remains:

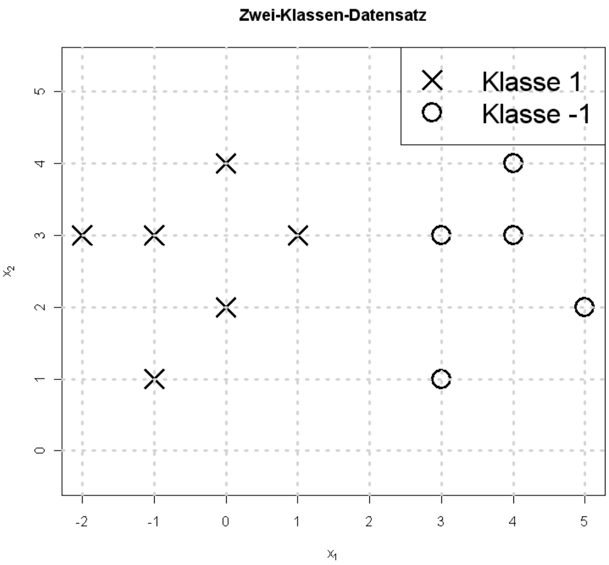
$$P(p \mid A=5,B=3) = \frac{P(A=5,B=3 \mid p)}{\int_0^1 P(A=5,B=3 \mid q)dq}$$

which we plug into the original formula:

$$\begin{aligned} E(\text{Bob wins}) &= \frac{\int_0^1 (1-p)^3 \cdot P(p \mid A=5,B=3)dp}{\int_0^1 P(A=5,B=3 \mid q)dq} \\ &\underset{\star}{=} \frac{\int_0^1 p^5(1-p)^6 dp}{\int_0^1 q^5(1-q)^3 dq} \\ &\underset{\text{Beta}}{=} \frac{5!6!}{12!}\frac{9!}{5!3!} \\ &= 1/11 \approx 0.09 \end{aligned}$$

Exercise 6-3 Optimal Seperating Hyperplane

Consider the following dataset consisting of points $\binom{x_1}{x_2}$ in $\mathbb{R}^2$. Using a hyperplane, points marked by × are to be mapped onto ≥ 1, points marked by ◯ are to be mapped onto ≤ -1.



(a) Find the support vectors.

(b) Determine the equation of the optimal separating hyperplane $h = \mathbf{x}^T\mathbf{w}$ and draw it within the figure.

(c) Compute the margin $\mathcal{C}$.

Exercise 6-4 Optimal Separating Hyperplane

Determine the optimal separating hyperplane of the following dataset, partitioned into two classes A and B:

$$A = \left\{ p_1 \; = \; \begin{pmatrix} 2 \\ 4 \end{pmatrix}, p_2 = \begin{pmatrix} 3 \\ -1 \end{pmatrix}, p_3 = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}, p_4 = \begin{pmatrix} 2.5 \\ 3 \end{pmatrix}, p_5 = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right\},$$

$$B = \left\{ p_6 = \begin{pmatrix} 0.5 \\ 1.5 \end{pmatrix}, p_7 = \begin{pmatrix} -1 \\ 3 \end{pmatrix}, p_4 = \begin{pmatrix} 0 \\ 0.5 \end{pmatrix} \right\}$$

Instances of class A shall be labeled with 1, instances of class B with -1.

Visualize the result and name the support vectors. How wide is the margin?

Possible Solution:

Solution by looking at the picture: $\{1,3,6\}$are the uspport vectors, thus, separating hyperplane is:

$$s_1 = \begin{pmatrix} 1.25 \\ 2.75 \end{pmatrix}, s_2 = \begin{pmatrix} 0.75 \\ 1 \end{pmatrix} \qquad \text{Geradengleichung: } h = 0 = \mathbf{w}_0 + \mathbf{w}_1 \cdot \mathbf{x}_1 + \mathbf{w}_2 \cdot \mathbf{x}_2$$

$$\begin{aligned} \Rightarrow \text{I)}: &\quad \mathbf{w}_0 + \mathbf{w}_1 \cdot 1.25 + \mathbf{w}_2 \cdot 2.75 = 0 \\ \text{II)}: &\quad \mathbf{w}_0 + \mathbf{w}_1 \cdot 0.75 + \mathbf{w}_2 \cdot 1 = 0 \\ \text{I)}-\text{II)}: &\quad \mathbf{w}_1 \cdot 0.5 + \mathbf{w}_2 \cdot 1.75 = 0 \qquad \Rightarrow \mathbf{w}_1 = -3.5 \cdot \mathbf{w}_2 \end{aligned}$$

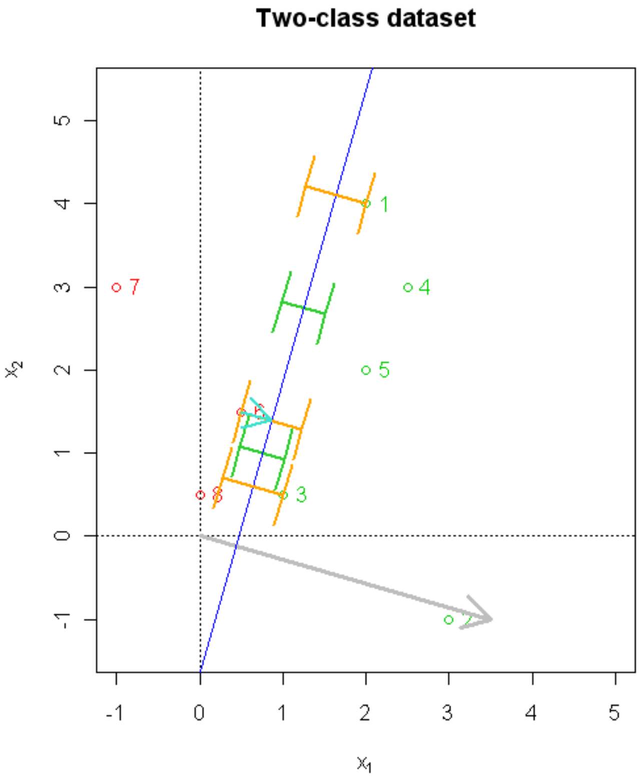
$$\text{definiere } \mathbf{w}_2 = 1 \Rightarrow \mathbf{w}_1 = -3.5 \overset{\text{in I)}}{\Rightarrow} \mathbf{w}_0 = 3.5 \cdot 1.25 - 1 \cdot 2.75 = 1.625$$

Margin condition: $y_i(\mathbf{x}_i^T \mathbf{w}_{\text{opt}}) \geq 1$. Test if correctly classified: $y_i \cdot \sum_{j=0}^{2} \mathbf{w}_j \mathbf{x}_{i,j} \overset{A_1}{=} 1 \cdot (1.625 - 3.5 \cdot 2 + 1 \cdot 4) = -1.375 \not\geq 1$. Hence, vectors needs to be negated: $(= \cdot(-1))$. $\Rightarrow \mathbf{w}^T = (-1.625, 3.5, -1)$. ($-1.625$ is the intersecting point with x_2-axis.)

Now we need the minimal $\mathbf{w}$. We have to minimize $\sum_{j=1}^{2} \mathbf{w}_j^2$. It is not valid to divide $\mathbf{w}$ by some sufficiently large number. e.g., 100, as this would violate the margin condition: $y_i \cdot \sum_{j=0}^{2} \mathbf{w}_j \mathbf{x}_{i,j} \geq 1$ for the first point. It is known that $y_i(\mathbf{x}_i^T \mathbf{w}_{\text{opt}}) = 1$ for the support vectors. In this case, $\min y_i(\mathbf{x}_i^T \mathbf{w}) = 1.375$, hence, it suffices, to divide $\mathbf{w}$ by 1.375 to find the optimal seperating hyperplane.

$$\mathbf{w}_{\text{opt}} = \frac{\mathbf{w}}{y_i(\mathbf{x}_i^T \mathbf{w}_{\text{opt}})} = \frac{\mathbf{w}}{1.375} = \begin{pmatrix} -1.\overline{18} \\ 2.\overline{54} \\ -0.\overline{72} \end{pmatrix}$$

The margin is defined as
$\mathcal{C} = \frac{1}{\|\tilde{\mathbf{w}}_{\text{opt}}\|} = \frac{1}{\|(\mathbf{w}_{\text{opt},1}, \mathbf{w}_{\text{opt},2})^T\|} = \frac{1}{\sqrt{3.5^2/1.375 + 1^2/1.375}} \approx \approx 1/3.1042 \approx 0.3777416.$


Two-class dataset
x2
x1