

Statistical Machine Learning

Summer Term 2021, Homework 1

Prof. Stefan Roth, Dr. Simone Schaub-Meyer



TECHNISCHE
UNIVERSITÄT
DARMSTADT

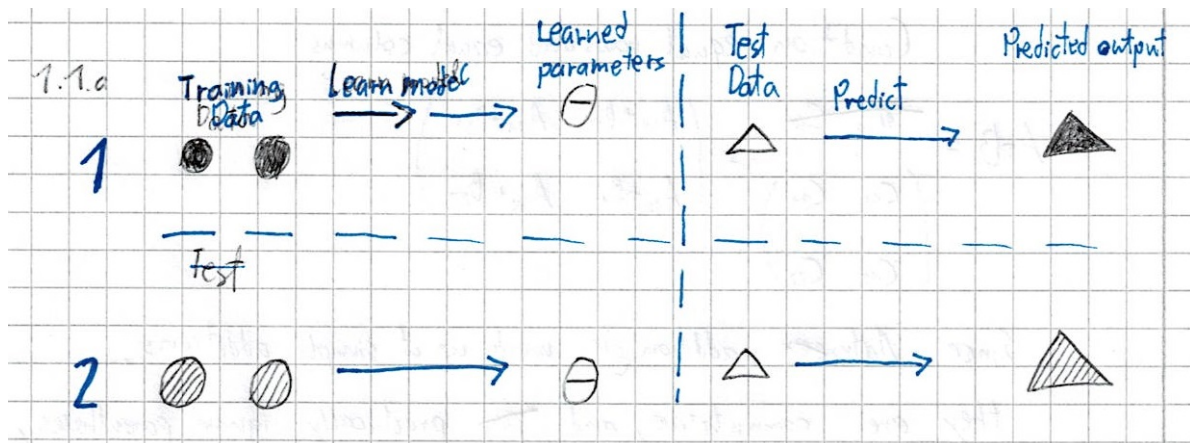
Total points: ?? + ?? bonus

Due date: 25. May 2018

Stefanie Martin, Maximilian Nothnagel

Problem 1.1 Machine Learning Introduction [6 Points]

a) Model Fitting [6 Points]



Model 1 is being trained on filled circles, and as such assumes that the triangle is also to be filled, coming to an incorrect result.

Model 2 is being trained on striped circles, and as such assumes that the triangle is also to be striped, coming to the correct conclusion.

Problem 1.2 Linear Algebra Refresher [20 Points]

a) Matrix Properties [5 Points]

1. Multiplication

$$A * B = \begin{pmatrix} A_{1,1} * B_{1,1} + A_{2,1} * B_{1,2} & A_{1,1} * B_{2,1} + A_{2,1} * B_{2,2} \\ A_{1,2} * B_{1,1} + A_{2,2} * B_{1,2} & A_{1,2} * B_{2,1} + A_{2,2} * B_{2,2} \end{pmatrix}$$

$A * B$ is defined only when Columns of B equal rows of A.

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} * \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$

BUT

$$\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} * \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Matrixmultiplication is not Commutative.

$$A(B + C) = AB + AC$$

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} * \left(\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 1 \\ 2 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 0 & 0 \end{pmatrix}$$

Matrixmultiplication is Distributive.

$$(A * B) * C = A * (B * C)$$

$$\left(\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} * \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} \right) * \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} * \left(\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right) = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$$

2. Addition Since matrix-addition is made up of simple additions, they are commutative and practically ignore parentheses, so are also distributive and associative.

The condition for any matrix-addition is that the matrices have both equal rows and columns.

$$A_{2,2} + B_{2,2} = \begin{pmatrix} C_{1,1} & C_{2,1} \\ C_{1,2} & C_{2,2} \end{pmatrix} = \begin{pmatrix} A_{1,1} + B_{1,1} & A_{2,1} + B_{2,1} \\ A_{1,2} + B_{1,2} & A_{2,2} + B_{2,2} \end{pmatrix}$$

b) Matrix Inversion [7 Points]

$$A^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

...Via Gauß-Jordan. Under the condition that $a = b = d = 0; c = 1$

$$\begin{pmatrix} 2 & 2 & 3 \\ 0 & 1 & 0 \\ 8 & 3 & 12 \end{pmatrix}$$

Is not invertable, since it's Determinant $\text{Det} = 0$.

c) Matrix Pseudoinverse [3 Points]

$$\text{Left: } A^\# * A = (A^T * A)^{-1} * A^T$$

$$\text{Right: } A * A^\# = A * A^T (A * A^T)^{-1}$$

Since $A_{2 \times 3}$ has more rows than columns, the left Moore-Penrose exists.

$$\text{The equation is: } A^\#_{3 \times 2} * A = (A^T_{3 \times 2} * A_{2 \times 3})^{-1}_{2 \times 2} * A^T_{3 \times 2}$$

d) Basis Transformation [5 Points]

Vector with new Basis $v^* = T^{-1} * v$

$$1) T_v = E^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; T_w = B^{-1} = \begin{pmatrix} 2 & -1.5 \\ -1 & 1 \end{pmatrix}$$

$$2) 2 * \begin{bmatrix} 2 \\ -1 \end{bmatrix} + 5 * \begin{bmatrix} -1 \\ 5 \end{bmatrix} = \begin{bmatrix} -3 \\ 5 \end{bmatrix} = v^*$$

Problem 1.3 Statistics Refreshner [0 Points]

a) Expectation & Variance [8 Points]

1. We can define the expectation by

$$E[f] = \sum_{w \in \Omega} P(w) f(w) \quad (18)$$

Which leads to the variance:

$$\text{var}[f] = E[f^2] - E[f]^2. \quad (19)$$

If we have 2 random variables X,Y and $Z = X + Y$. Then is the expectation a linear function, since for any 2 points

$$E[Z] = \sum_{w \in \Omega} Z(w) P(w) = \sum_{w \in \Omega} (X(w) + Y(w)) P(w) = E[X] + E[Y] \quad (20)$$

applies. Since the variance of the sum of 2 random variables is

$$\text{var}[Z] = E[Z^2] - E[Z]^2 = \text{var}[X] + \text{var}[Y] + 2E[XY] - 2E[X]E[Y] \text{ is } E[XY] \neq E[X]E[Y] \quad (21)$$

$$\text{And that is not a linear operator.} \quad (22)$$

2. Unbiased estimator:

$$\bar{x} = \frac{1}{n} * \sum_{i=1}^n x_i \quad (23)$$

$$\overline{xA} = \frac{1}{6} * (1 + 5 + 6 + 3 + 2 + 1) = 3 \quad (24)$$

$$\overline{xB} = \frac{1}{6} * (6 + 1 + 1 + 4 + 1 + 5) = 3 \quad (25)$$

$$\overline{x\bar{C}} = \frac{1}{6} * (3 + 2 + 3 + 3 + 4 + 3) = 3 \quad (26)$$

unbiased estimator for the variance

$$\overline{\sigma} = \frac{1}{n-1} * \sum_{i=1}^n x_i - \bar{x}^2 \quad (27)$$

$$\overline{\sigma A} = \frac{1}{5} * ((1-3)^2 + (5-3)^2 + (6-3)^2 + (3-3)^2 + (2-3)^2 + (1-3)^2) = \frac{22}{5} = 4,4 \quad (28)$$

$$\overline{\sigma B} = \frac{1}{5} * ((6-3)^2 + (1-3)^2 + (1-3)^2 + (4-3)^2 + (1-3)^2 + (5-3)^2) = \frac{26}{5} = 5,2 \quad (29)$$

$$\overline{\sigma \bar{C}} = \frac{1}{5} * ((3-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (3-3)^2 + (3-3)^2) = \frac{2}{5} = 0,4 \quad (30)$$

3.

$$KL : \sum_{x \in X} P(x) \ln \frac{P(x)}{Q(x)} \quad (31)$$

$$KL(PA \parallel Q) = \frac{3}{6} * \ln\left(\frac{\frac{3}{6}}{\frac{1}{6}}\right) + \frac{1}{6} * \ln\left(\frac{\frac{1}{6}}{\frac{1}{6}}\right) + \frac{1}{6} * \ln\left(\frac{\frac{1}{6}}{\frac{1}{6}}\right) + \frac{1}{6} * \ln\left(\frac{\frac{1}{6}}{\frac{1}{6}}\right) + \frac{1}{6} * \ln\left(\frac{\frac{1}{6}}{\frac{1}{6}}\right) = 3 \quad (32)$$

$$KL(PB \parallel Q) = \frac{1}{6} * \ln\left(\frac{\frac{1}{6}}{\frac{1}{6}}\right) + \frac{2}{6} * \ln\left(\frac{\frac{2}{6}}{\frac{1}{6}}\right) + \frac{1}{6} * \ln\left(\frac{\frac{1}{6}}{\frac{1}{6}}\right) + \frac{1}{6} * \ln\left(\frac{\frac{1}{6}}{\frac{1}{6}}\right) = 1,52 \quad (33)$$

$$KL(PC \parallel Q) = \frac{4}{6} * \ln\left(\frac{\frac{4}{6}}{\frac{1}{6}}\right) + \frac{1}{6} * \ln\left(\frac{\frac{1}{6}}{\frac{1}{6}}\right) + \frac{1}{6} * \ln\left(\frac{\frac{1}{6}}{\frac{1}{6}}\right) = 2.38 \quad (34)$$

A has the biggest KL divergence, so it is the closest

b) It is a cold world [7 Points]

1.

$$a \in \{0, 1\} : \text{if a person has backspin, with } 1 = \text{pain and } 0 = \text{no pain} \quad (43)$$

$$b \in \{0, 1\} : \text{if a person has a cold, with } 1 = \text{pain and } 0 = \text{no cold} \quad (44)$$

2.

$$P(a = 1 \mid b = 1) = 0,25 \quad (45)$$

$$P(b = 1) = 0,04 \quad (46)$$

$$P(a = 1 \mid b = 0) = 0,1 \quad (47)$$

3. Rule of Bayes

$$P(b = 1 \mid a = 1) = \frac{P(a = 1 \mid b = 1)P(b = 1)}{P(b = 1)} \quad (48)$$

$$\frac{P(a = 1 \mid b = 1)P(b = 1)}{P(a = 1 \mid b = 1)P(b = 1) + P(a = 1 \mid b = 0)P(b = 0)} \quad (49)$$

$$\text{Werte einsetzen: } \frac{0,25 * 0,04}{0,25 * 0,04 + 0,10 * (1 - 0,04)} = \frac{5}{53} \approx 0,094 \quad (50)$$

c) Cure the virus [14 Points]

1. Markov Chain

$$S_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (62)$$

$$S_1 = \begin{pmatrix} 0,42 \\ 0,58 \end{pmatrix} \quad (63)$$

$$P = \begin{pmatrix} 0,42 & 0,974 \\ 0,58 & 0,026 \end{pmatrix} \quad (64)$$

$$S_1 = P * S_0 \leftrightarrow \begin{pmatrix} 0,42 \\ 0,58 \end{pmatrix} = \begin{pmatrix} 0,42 & 0,974 \\ 0,58 & 0,026 \end{pmatrix} * \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad (65)$$

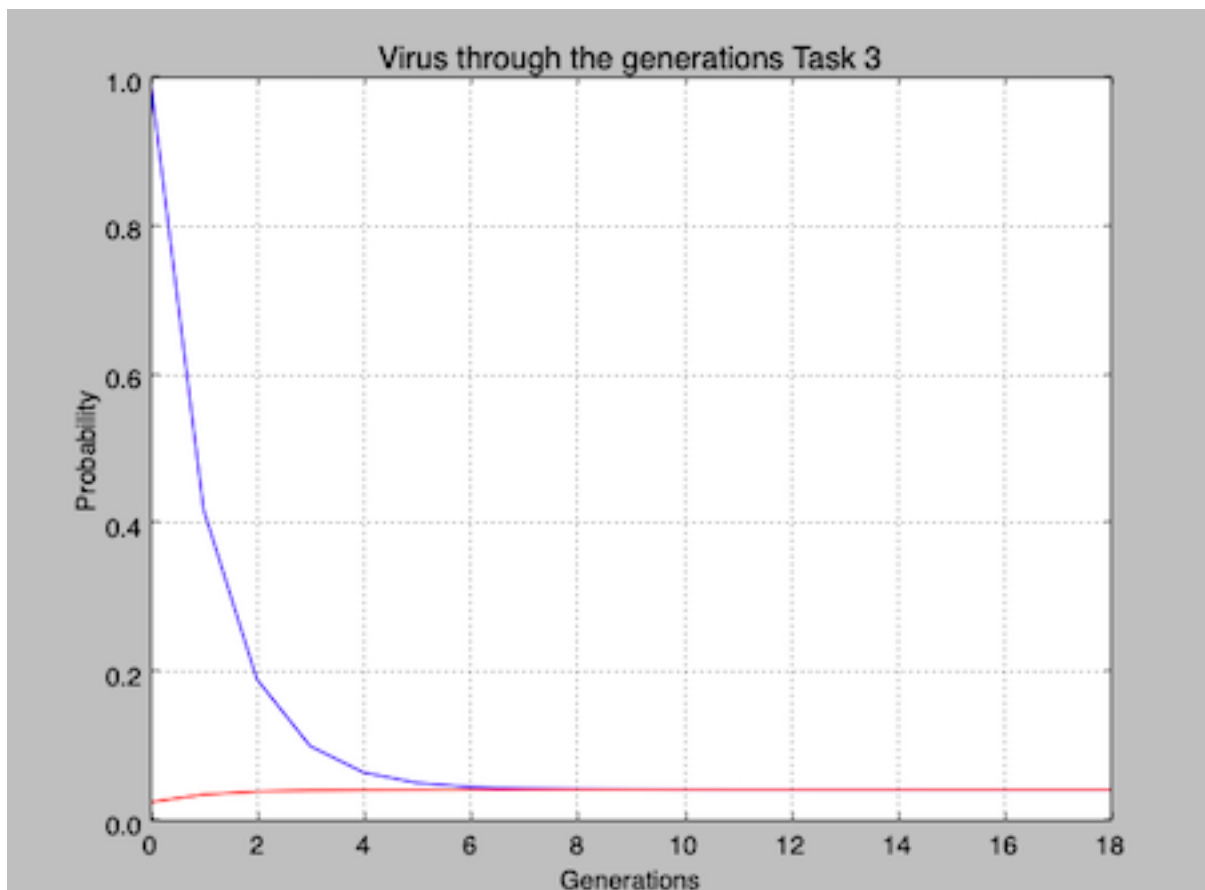
$$S_2 = P * S_1 \leftrightarrow \begin{pmatrix} 0,741 \\ 0,394 \end{pmatrix} = \begin{pmatrix} 0,42 & 0,974 \\ 0,58 & 0,026 \end{pmatrix} * \begin{pmatrix} 0,42 \\ 0,58 \end{pmatrix} \quad (66)$$

2.

```
def markovChain(s0, p, g):
    s = s0.copy()
    result = np.zeros(g+1)
    result[0]=s[0]
    for i in range(1, g+1):
        s = s.dot(p)
        result[i]=s[0]
    return result

s0 = np.array([1, 0])
s1 = np.array([0.026, 0.974])
p = np.array([[0.42, 0.58],[0.026, 0.974]])
n = np.arange(0, 19, 1)

gen18 = markovChain(s0, p, 18)
gen18prog = markovChain(s1, p, 18)
```



3. After 6 timesteps does the ratios stop to change significantly. Stable probability:

$$P = \begin{pmatrix} 0.42 & 0.974 \\ 0.58 & 0.026 \end{pmatrix} \quad (67)$$

$$\bar{X} = \begin{pmatrix} A \\ B \end{pmatrix} \quad (68)$$

$$P * \bar{X} = \bar{X} \leftrightarrow \begin{pmatrix} 0,42 & 0,974 \\ 0,58 & 0,026 \end{pmatrix} * \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} A \\ B \end{pmatrix} \quad (69)$$

$$0,42A + 0,5B = A \Rightarrow 0,5B = A - 0,42A \Rightarrow B = \frac{25}{29} = 0,86 \quad (70)$$

$$0,58A + 0,026B = B \quad (71)$$

$$A + B = 1 \Rightarrow A + 0,86 = 1 \Rightarrow A = 0,14 \quad (72)$$

We can see the probability converge to our solution.

Problem 1.4 Information Theory [0 Points]

a) Entropy [5 Points]

1.

$$-0,04 * \log_2 0,04 - 0,22 * \log_2 0,22 - 0,67 * \log_2 0,67 - 0,07 * \log_2 0,07 = 1,3219 \quad (75)$$

An average of 1 bit can be transmitted.

2.

$$H = \ln(4) = 1,386 \approx 2 \quad (76)$$

Maximum of 2 bits per symbol can be transmitted using a set of four symbols. The distribution over the symbols requires that at least the maximum is as great as that of all other members.

Problem 1.5 Bayesian Decision Theory [0 Points]

a) Optimal Boundary [4 Points]

1. Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification based probabilities.
2. The goal is to decide which class an example x most likely belongs to. This is done by comparing the class posterior probabilities $p(C_i | x)$, which can be calculated by Bayes' theorem:

$$p(C_i | x) = \frac{p(x | C_i)p(C_i)}{p(x)} \propto p(x | C_i)p(C_i)$$

3. The decision boundary of two classes C_1 and C_2 is given by $p(C_1 | x) = p(C_2 | x)$, where C_1 is chosen over C_2 if $p(C_1 | x) > p(C_2 | x)$.

b) Decision Boundaries [8 Points]

c) Different Misclassification Cost [8 Points]

- 1.
- 2.