Total points: ?? + ?? bonusDue date: 25. May 2018

Statistical Machine Learning

Summer Term 2021, Homework 1

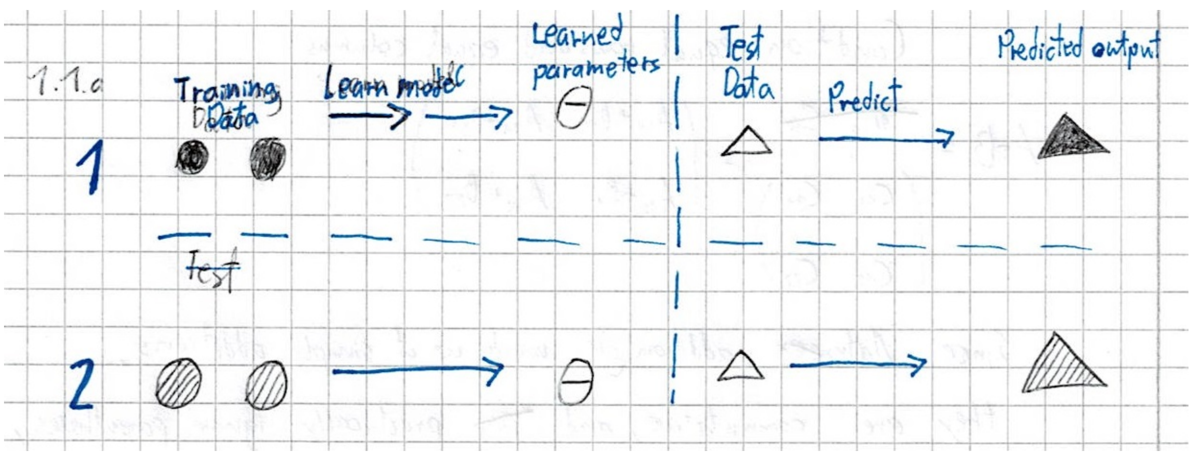
Prof. Stefan Roth, Dr. Simone Schaub-Meyer

May 25, 2021

Stefanie Martin, Maximilian Nothnagel
Stefanie Martin, Maximilian Nothnagel

1.1 Machine Learning Introduction [0 Points]

1. **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.

1.2 Linear Algebra Refresher [0 Points]

1. **Matrix Properties [5 Points]**

(a) *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.

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

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

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

(b) *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.

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

2. **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 $Det = 0$.

3. **Matrix Pseudoinverse [3 Points]**

Left: $A^{\#} * A = \left(A^T * A\right)^{-1} * A^T$
Right: $A * A^{\#} = A * A^T \left(A * A^T\right)^{-1}$
Since A_{2x3} *has more rows than columns, the left Moore-Penrose exists.*
The equation is: $A^{\#}_{3x2} * A = \left(A^T_{3x2} * A_{2x3}\right)^{-1}_{2x2} * A^T_{3x2}$

4. **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) $2 * \begin{bmatrix} 2 \\ -1 \end{bmatrix} + 5 * \begin{bmatrix} -1,5 \\ 1 \end{bmatrix} = \begin{bmatrix} -3,5 \\ 3 \end{bmatrix} = v^*$

1.3 Statistics Refreshner [0 Points]

1. **Expectation & Variance [8 Points]**

   (a) *We can define the expectation by*

   $$E|f| = \sum_{w\in\Omega} P(w) f(w) \tag{18}$$

   Which leads to the variance:

   $$var[f] = E\left|f^2\right| - E[f]^2. \tag{19}$$

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

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

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

   $$var[Z] = E\left[Z^2\right]-E[Z]^2 = var|X|+var[Y]+2E[XY]-2E[X]E[Y] isE[XY] \neq E[X]E[Y] \tag{21}$$

   $$Andthatisnotalinearoperator. \tag{22}$$

   (b) *Unbiased estimator:*

   $$\overline{x} = \frac{1}{n} * \sum_{i=1}^{n} x_i \tag{23}$$

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

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

   $$\overline{xC} = \frac{1}{6} * (3+2+3+3+4+3) = 3 \tag{26}$$

   unbiased estimator for the variance

   $$\overline{\sigma} = \frac{1}{n-1} * \sum_{i=1}^{n} x_i - \overline{x}^2 \tag{27}$$

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

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

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

(c)

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

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

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

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

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

2. **It is a cold world [7 Points]**

(a)

$$a \in \{0,1\} : ifapersonhasbackspin, with 1 = painand 0 = nopain \tag{43}$$

$$b \in \{0,1\} : ifapersonhasacold, with 1 = painand 0 = nocold \tag{44}$$

(b)

$$P(a = 1 \mid b = 1) = 0,25 \tag{45}$$

$$P(b = 1) = 0,04 \tag{46}$$

$$P(a = 1 \mid b = 0) = 0,1 \tag{47}$$

(c) *Rule of Bayes*

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

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

$$Werteeinsetzen : \frac{0,25 * 0,04}{0,25 * 0,04 + 0,10 * (1 - 0,04)} = \frac{5}{53} \approx 0,094 \tag{50}$$

3. **Cure the virus [14 Points]**

(a) *Markov Chain*

$$S_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{62}$$

$$S_1 = \begin{pmatrix} 0,42 \\ 0,58 \end{pmatrix} \tag{63}$$

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

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

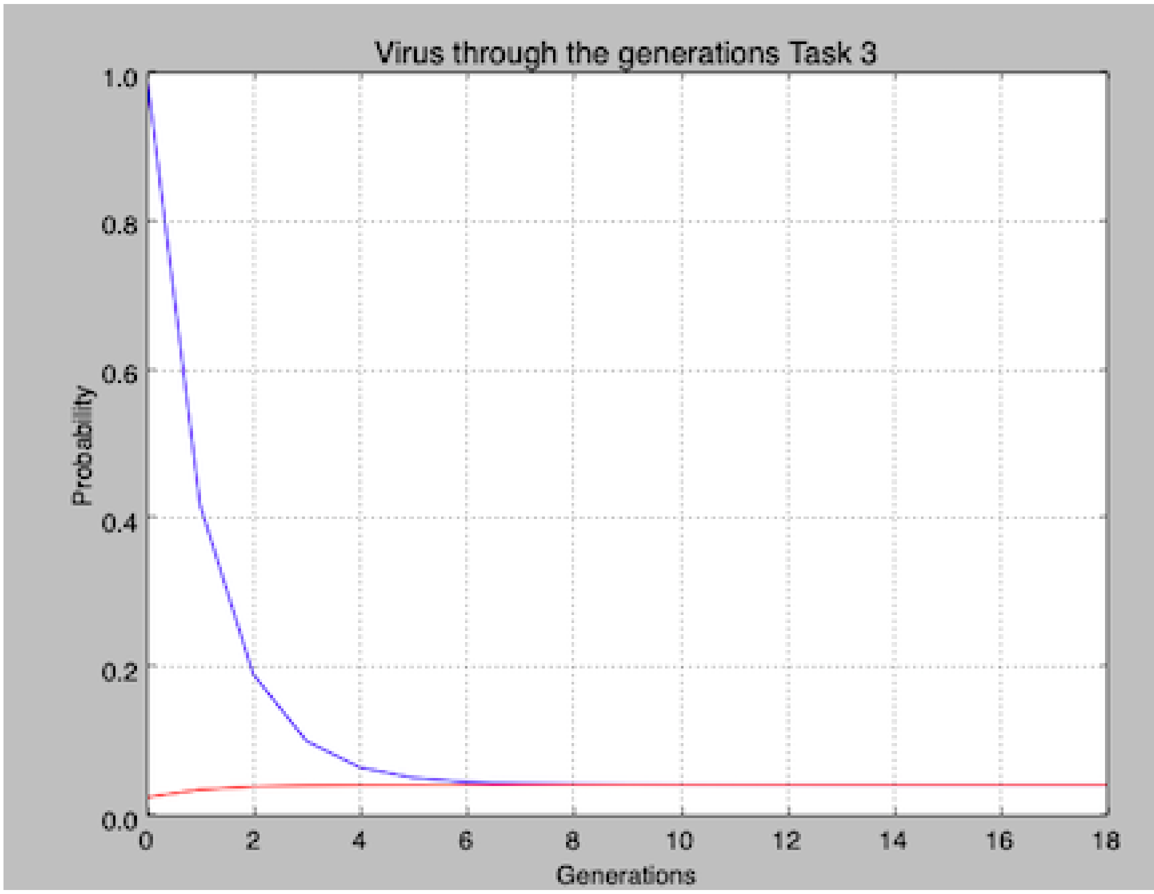
$$S_2 = P * S_1 \leftrightarrow \begin{pmatrix} 0,741 \\ 0,394 \end{pmatrix} = \begin{pmatrix} 0,42 & 0,974 \\ 0,58 & 0,026 \end{pmatrix} * \begin{pmatrix} 0,42 \\ 0,58 \end{pmatrix} \tag{66}$$

(b)

```
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)
```



(c) *After 6 timesteps does the ratios stop to change significantly. Stable probability:*

$$P = \begin{pmatrix} 0,42 & 0,974 \\ 0,58 & 0,026 \end{pmatrix} \tag{67}$$

$$\overline{X} = \begin{pmatrix} A \\ B \end{pmatrix} \tag{68}$$

$$P * \overline{X} = \overline{X} \leftrightarrow \begin{pmatrix} 0,42 & 0,974 \\ 0,58 & 0,026 \end{pmatrix} * \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} A \\ B \end{pmatrix} \tag{69}$$

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

$$0,58A + 0,026B = B \tag{71}$$

$$A + B = 1 \Rightarrow A + 0,86 = 1 \Rightarrow A = 0,14 \tag{72}$$

We can see the probability converge to our solution.

1.4 Information Theory [0 Points]

1. **Entropy [5 Points]**

   (a)

   $$-0,04*\log_2 0,04-0,22*\log_2 0,22-0,67*\log_2 0,67-0,07*\log_2 0,07 = 1,3219 \tag{75}$$

   An average of 1 bit can be transmitted.

   (b)

   $$H = ln\,(4) = 1,386 \approx 2 \tag{76}$$

   Maxmimum 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.

1.5 Bayesian Decision Theory [0 Points]

1. **Optimal Boundary [4 Points]**

   (a) *Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification based probabilities.*

   (b) *The goal is to decide which class an example x most likely belongs to. This is done by comparing the class posterior probabilities $p\,(C_i \mid x)$, which can be calculated by Bayes' theorem:*

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

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

2. **Decision Boundaries [8 Points]**

3. **Different Misclassification Cost [8 Points]**

   (a)

   (b)