Fatih Beşker
150210710

Q1.1) In real life, it is hard to explain how certain actions are done. For example, we don't know how do we differentiate object. We intuitively differentiate them but we can't tell how. Or some of the problems are beyond of our understanding, such as customer actions in a market place. In both cases, there are no clear instructions for computers to calculate things. In the cases like this, machine learning is used.

Machine learning is the process of computers can learn and adapt without clear instructions. To achieve this, large amount of data is analyzed using some algorithms and statistical inferences. This is the "learning" part of the machine learning.

Machine learning can be divided into three categories. These are supervised learning, unsupervised learning and reinforcement learning.

Supervised learning is the learning where the sample inputs are given and the sample outputs are labeled. Label can be continuous or discrete. For example, we have some samples from a factory line and certain measurements are taken from them. Experts label the samples as faulty or not-faulty. Given this data, it is possible to train a model which would be used for classifying products as faulty or not-faulty. This is an example for supervised learning.

In the unsupervised learning, the data is not labeled, and the machine is tasked for finding "clusters" of data which might resemble a class. This is called "clustering". Unsupervised learning is used for pattern recognition, anomaly detection, recommendation engines and such. In each example, the data is not labeled and machine tries to find classes in the data.

In reinforcement learning, the desired behaviors are rewarded and undesired ones are punished. The result comes after the action and machine updates its actions according to the previous results. It is a learning process for decision making, machine learns to make optimal decisions trough trial and error. AIs made for beating certain games (chess AI) or self driving cars learn this way.

Fatih Baskın
150210710.

Q 1.2

From a given data sample, large portion of it is used for training a model, the remaining part is used for testing the model.

Let's say a model is trained with a training set. This is the training phase. Then, this model will set its parameters optimally for the test data sample, to be explicit, it yields the least amount of error possible for the given training data sample. But, for this model to have good generalization characteristic, it should return good answers for any arbitrary input data. Therefore, its generalization performance should be tested. Test data is used for this, and if the test performance is good, then it is likely that the trained model will have good generalization performance.

Q1.3)

The linear regression is used for supervised machine learning. The labeled input-output samples are provided to the learning process and in the end, model will try to predict outputs with least possible error.

Let's say there are features $x_1^{(i)}, x_2^{(i)} \dots x_d^{(i)}$ and output $y^{(i)}$ for a given data sample. Machine will have a parametric function which uses the input features and some corresponding coefficients θ to make a prediction $\hat{y}$. The process of linear regression will yield θ values of which the prediction function (hypothesis) will yield the minimum error.

$$\underbrace{h_\theta(x^{(i)})}_{\text{hypothesis function}} = x^{(i)}\theta = \hat{y}^{(i)} \rightarrow \frac{1}{N}\sum_{i=1}^{N}(\underbrace{y^{(i)}}_{\text{actual outcome}} - \underbrace{\hat{y}^{(i)}}_{\text{prediction}})^2 = \frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - x^{(i)}\theta) \rightarrow \text{Mean square error.}$$

$$= J(\theta) \rightarrow \text{objective function.}$$

Aim of the linear regression is to make accurate predictions about the outcomes given the features. It is quite useful in finance where it allows to make predictions about a company given company financial data. Other usecase is the marketing, where with the regression, it is possible to make predictions about expected sales, etc.

Fatih Basker
150210710

Q1.4)

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$J(\theta) = \sum_{i=1}^{7} (y^{(i)} - \hat{y}^{(i)})^2 = \sum_{i=1}^{7} (y - \theta_0 - \theta_1 x_1)^2 \rightarrow$$ objective function

aim is to minimize the square error.

θ vector, $\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$

$$\sum_{i=1}^{7} (y - \theta_0 - \theta_1 x_1)^2 = (Y - X\theta)^T (Y - X\theta) = Y^T Y - Y^T X\theta - \theta^T X^T Y - \theta^T X^T X \theta$$

local minima $\rightarrow$ minimum of squared error is at the derivative, $\frac{\partial J(\theta)}{\partial \theta} = 0$

$$0 = \frac{\partial}{\partial \theta} \left(Y^T Y - Y^T X \theta - \theta^T X^T Y - \theta^T X^T X \theta \right) = \frac{\partial J(\theta)}{\partial \theta}$$

$$0 = \frac{\partial}{\partial \theta} Y^T Y - \frac{\partial}{\partial \theta} (y^T X \theta) - \partial(\theta^T X^T y) - \frac{\partial}{\partial T} (\theta^T X^T X \theta)$$

$$= 0 - X^T Y - X^T Y - 2 X^T X \theta = -2 X^T Y - 2 X^T X \theta$$

$$X^T X \theta = X^T Y \rightarrow \theta = (X^T X)^{-1} X^T Y$$

x, y tuples, $\hat{y} = \theta_0 + \theta_1 x$,
$= \underline{1} \cdot \theta_0 + \theta_1 . \underline{x}$

$$X = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \quad Y = \begin{bmatrix} 5 \\ 8 \\ 7 \\ 10 \\ 12 \\ 14 \\ 15 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} = \begin{bmatrix} 7 & 35 \\ 35 & 203 \end{bmatrix}$$

$$[I \mid X^T X] = \left[\begin{array}{cc|cc} 1 & 0 & 7 & 35 \\ 0 & 1 & 35 & 203 \end{array}\right]$$

$\downarrow$, $r_1 = r_1 / 7$

$$\left[\begin{array}{cc|cc} 1/7 & 0 & 1 & 5 \\ 0 & 1 & 35 & 203 \end{array}\right]$$

$\downarrow$ $r_2 = r_2 - 35 r_1$

$$\left[\begin{array}{cc|cc} 1/7 & 0 & 1 & 5 \\ -5 & 1 & 0 & 28 \end{array}\right]$$

$\downarrow$, $r_2 = \frac{r_2}{28}$

$$\left[\begin{array}{cc|cc} 1/7 & 0 & 1 & 5 \\ \frac{-5}{28} & \frac{1}{28} & 0 & 1 \end{array}\right] \xrightarrow{r_1 = r_1 - 5 r_2} \underbrace{\left[\begin{array}{cc|cc} \frac{29}{28} & \frac{-5}{28} & 1 & 0 \\ \frac{-5}{28} & \frac{1}{28} & 0 & 1 \end{array}\right]}_{(X^T X)^{-1} \mid I}$$

$$\theta = (X^T X)^{-1} X^T Y = \begin{bmatrix} 29/28 & -5/28 \\ -5/28 & 1/28 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{bmatrix} \begin{bmatrix} 5 \\ 8 \\ 7 \\ 10 \\ 12 \\ 14 \\ 15 \end{bmatrix}$$

$$\theta = \begin{bmatrix} 29/28 & -5/28 \\ -5/28 & 1/28 \end{bmatrix} \begin{bmatrix} 71 \\ 402 \end{bmatrix}$$

$$\theta = \begin{bmatrix} 7/4 \\ 47/28 \end{bmatrix} \rightarrow$$ So,

$$\hat{y} = \underbrace{\frac{7}{4}}_{\theta_0} + \underbrace{\frac{47}{28}}_{\theta_1} x$$

$$\boxed{\begin{array}{l} \text{Slope: } 47/28 \\ \text{y-Intercept: } 7/4 \end{array}}$$

Fatih Baskın
150210710

Q 2.1) Maximum likelihood estimation is the process of estimating the parameters of a probability distribution. For the given data, parameters of the probability distribution are estimated as the values which yield the highest probability for the given data.

$P(x \setminus \theta) \rightarrow$ likelihood, Likelihood of the x data given the θ parameters.

In MLE, the θ values that yield the highest $P(x \setminus \theta)$ values are estimated.

Q 2.2) $P(x_{1:n} \setminus \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)$

log of this : $L(\mu,\sigma) = \sum_{i=1}^{n}\left(\log\left(\frac{1}{\sqrt{2\cdot\pi\sigma^2}}\right) + \frac{-(x_i-\mu)^2}{2\sigma^2}\right)$

$= n\log\left(\frac{1}{\sqrt{2\pi}\,\sigma}\right) + \frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)$

$\frac{\partial L(\mu,\sigma)}{\partial \mu} = 0 + \frac{-1}{2\sigma^2}\left(\sum_{i=1}^{n}\left(2\cdot(x_i-\mu)\cdot(-1)\right)\right) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{N}(x_i-\mu)\right)$

$= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}x_i - n\cdot\mu\right) = 0 \Rightarrow \sum_{i=1}^{n}x_i = n\cdot\mu,\ \boxed{\mu_{ML} = \frac{\sum_{i=1}^{n}x_i}{n}}$

$\frac{\partial L(\mu,\sigma)}{\partial \sigma} = \frac{\partial}{\partial\sigma}\left(\left(n\log\left(\frac{\overbrace{\boxed{1}}^{B}}{\boxed{\sqrt{2\pi}}\,\sigma}\right)\right) + \overbrace{\boxed{\left(\frac{-1\left(\sum_{i=1}^{n}(x_i-\mu)^2\right)}{2}\right)}}^{A}\frac{1}{\sigma^2}\right)$

$= \frac{\partial}{\partial\sigma}\left(n\log(B\cdot\sigma^{-1}) + A\sigma^{-2}\right) = \frac{n\cdot B\cdot -\sigma^{-2}}{B\cdot\sigma^{-1}} + A\cdot(-2)\sigma^{-3} = -n\cdot\sigma^{-1} - 2A\sigma^{-3} = 0$

$-\sigma^{-1}(n + 2A\sigma^{-2}) = 0$, first root : $\sigma = 0$, second root : $n + 2A\sigma^{-2} = 0$, $-2A\sigma^{-2} = n$, $\frac{-2A}{n} = \sigma^2$

$\frac{\cancel{-2}\,\cancel{(-1)}\sum_{i=1}^{n}(x_i-\mu)^2}{\cancel{2}\cdot n} = \sigma^2$, $\boxed{\sigma^2_{ML} = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$

The given dataset : 6,33397611, 5,05162616, 3,44476117, 6,89726764, 6,40334359

$\mu_{ML} = \frac{\sum_{i=1}^{5}x_i}{5} = 5,62619494$ $\sigma^2_{ML} = \frac{\sum_{i=1}^{5}(x_i - 5,62619494)^2}{5} = 1.56186452$

$\sigma_{ML} = \sqrt{\sigma^2} = \sqrt{1.56186452} = 1.24974578$

The predictions μ_{ML} and σ_{ML} are different than the actual μ and σ which are used to create data samples. It is the downside of the maximum likelihood estimation since it is limited only to the small data sample that we observe.

Fatih Baskın
150210710

Q3.1)

Bayesian desicion theory combines probability and desicion making. Using the Bayes rule, probability of the each choice (or class) is evaluated and desicions on a given input is made using the probabilities of each class (choice).

$$\underbrace{P(C\backslash x)}_{} = \frac{\overbrace{P(x\backslash C)}^{\text{likelihood}}\ \overbrace{P(C)}^{\text{prior}}}{\underbrace{P(x)}_{\text{evidence}}}$$

- Posterior probability: Given the input parameters, probability of the class being C.
- Likelihood: Given a probabilistic distribution of a class, likelihood of the input parameters occuring in this distribution.
- Prior: The prior knowledge about a class or choice. for example, 60% of the objects are of type C_i.
- Evidence: Probability of the input parameters occuring. Marginalization of the input parameters is used to calculate this, meaning likelihood of all input parameters x is summed for all classes.

$$P(x) = \sum_C P(x\backslash C)\, P(C)$$

- It is also possible to combine the posterior probability with a loss function, where each desicion has a weight and with the probabilities, the desicion with the least harm (or loss) is decided. Weight and probabilities are multiplied to calculate the loss.

	$C=1$	$C=0$
$P(C=1\backslash x)$	l_1	l_2
$P(C=0\backslash x)$	l_3	l_4

Desicion is made according to the loss of each choice:

If we chose 1:

$$loss_{C=1} = l_1 \cdot P(C=1\backslash x) + l_3 \cdot P(C=0\backslash x)$$

If we choose 0:

$$loss_{C=0} = l_2 \cdot P(C=0\backslash x) + l_4 \cdot P(C=0\backslash x)$$

Desicion is made by the calculated losses of the desicions, therefore the desicion with the least loss is chosen.

Ex: Let's say we train a model in a factory line doing quality assesment. Deciding faulty product as pass is a very bad desicion so the loss of this option could be very high to prevent faulty products to pass from the quality assesment.

Fatih Baskın
150210710

Q3.2: X: test; can be positive or negative (1,0)
C: patient; can be sick or healthy (1,0)

- Sensitivity: Probability of a positive test given the patient is sick.
  $P(x=1 \backslash C=1) = 0{,}95$; $P(x=0 \backslash C=1) = 1 - P(x=1 \backslash C=1) = 1 - 0{,}95 = 0{,}05$
- Specifity: Probability of a negative test given the patient is healthy
  $P(x=0 \backslash C=0) = 0{,}90$; $P(x=1 \backslash C=0) = 1 - P(x=0 \backslash C=0) = 1 - 0{,}90 = 0{,}10$
- Prevalence: Probability of a person being sick.
  $P(C=1) = 0{,}05$; $P(C=0) = 1 - P(C=1) = 1 - 0{,}05 = 0{,}95$

$$\underbrace{P(C=1 \backslash x=1)}_{\text{posterior}} = \frac{\overbrace{P(x=1 \backslash C=1)}^{\text{likelihood}} \overbrace{P(C=1)}^{\text{prior}}}{\underbrace{P(x=1)}_{\text{evidence}}} = \frac{P(x=1 \backslash C=1) P(C=1)}{P(x=1 \backslash C=1) P(C=1) + P(x=1 \backslash C=0) P(C=0)}$$

- posterior: Given the test is positive, probability of patient being sick.
- likelihood: Given the patient is sick, probability of test being positive.
- prior: Probability of patient being sick, without any conditionality.
- evidence: Probability of test being positive.

$$P(C=1 \backslash x=1) = \frac{\cancel{0{,}95} \times 0{,}05}{\cancel{0{,}95} \times 0{,}05 + 0{,}10 \times \cancel{0{,}95}} = \frac{0{,}05}{0{,}05 + 0{,}10} = \frac{1}{3} = 0{,}\overline{33}$$