Vraag 1

Beantwoord op: 9 november 2021 - 08:45 **Duur:** 24 sec. **Score:** 1 van 1 pt.

1 pt.

When making a classication decision based on Bayes' law, the evidence (denominator) does not need to be computed for each class, because:

- It does not affect the probability of the most likely class
- It cannot change the probabilities so much that the classication decision would be changed
- You do need to compute it, actually, because otherwise you're not comparing probabilities
- ✓ It is the same for all classes 1 pt.

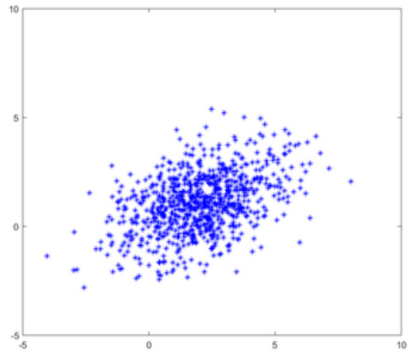
1 pt.

Vraag 2

Beantwoord op: 9 november 2021 - 08:51 **Duur:** 7 min. en 11 sec. **Score:** 0 van 1 pt.

0 pt.

Consider the following visualization of a two dimensional dataset.



Select the angle which will capture maximum variability along a single axis (i.e. the angle made by the first principal component with Ox).

- ✗ ~ 90 degree
- ~ 0 degree
- ~ 45 degree 1 pt.
- ~ 120 degree

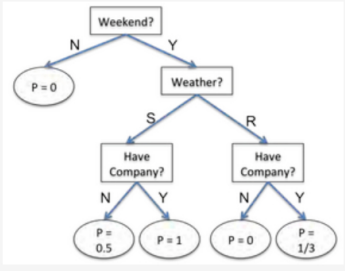
0 pt.

Vraag 3

Beantwoord op: 9 november 2021 - 08:54 **Duur:** 3 min. en 11 sec. **Score:** 1 van 1 pt.

1 pt.

We have some data about when people go hiking. The data take into effect, wether hike is on a weekend or not, if the weather is rainy (R) or sunny (S), and if the person will have company during the hike. Base on this we can drow the following decision tree.



According to the decision tree, what is the probability of going to hike on a rainy week day, without any company?; How about probability of going to hike on a rainy weekend when having some company?

In the decision tree, we used the following notations: no(N), yes(Y), rainy(R), sunny(S), and probability(P).

- ✓ the probability of going to hike on a rainy week day, without any company is 0; and the probability of going to hike on a rainy weekend when having some company is 1/3. 1 pt.
- the probability of going to hike on a rainy week day, without any company is 0.5; and the probability of going to hike on a rainy weekend when having some company is 1/3.
- the probability of going to hike on a rainy week day, without any company is 0.5; and the probability of going to hike on a rainy weekend when having some company is 1.
- the probability of going to hike on a rainy week day, without any company is 0; and the probability of going to hike on a rainy weekend when having some company is 1.

Vraag 4 **Beantwoord op:** 9 november 2021 - 08:54 **Duur:** 39 sec. **Score:** 1 van 1 pt. **1 pt.**

You are using a classification technique to a dataset, and you get a bad performance on the training set, but a good performance on the test set. How is this possible?

- ☐ The classifier is not well suited for this problem
- ☐ ☐ The test set is way too small to be representative 1 pt. ☐
- ☐ The training set is way too small for the model to train properly
- ☐ The classifier is overtting on the training set

1 pt.

Vraag 5 **Beantwoord op:** 9 november 2021 - 09:00 **Duur:** 5 min. en 52 sec. **Score:** 1 van 1 pt. **1 pt.**

The effectiveness of an SVM depends upon:

- ☐ Soft Margin Parameter C
- ☐ ☐ All three statements are true. 1 pt. ☐
- ☐ Selection of Kernel
- ☐ Kernel Parameters

1 pt.

Vraag 6 **Beantwoord op:** 9 november 2021 - 09:00 **Duur:** 25 sec. **Score:** 1 van 1 pt. **1 pt.**

Consider the following confusion matrix

		Predicted class		
		C_1	C_2	C_3
Actual Class	C_1	110	8	7
	C_2	16	130	10
	C_3	26	5	120

Given the above confusion matrix. What is the recall for class C_1?

- ☐ 110/152
- ☐ ☐ 110/125 1 pt. ☐
- ☐ 110/172
- ☐ 110/432

1 pt.

Vraag 7 **Beantwoord op:** 9 november 2021 - 09:01 **Duur:** 1 min. en 48 sec. **Score:** 0 van 1 pt. **0 pt.**

In Machine Learning, when we talk about a linear classifier, we consider a classification function that:

- ☐ Is a linear function of the input data, but not necessarily of the parameters
- ☐ ☐ Is a linear function of the data and parameters, and therefore separates the input data into two different classes using a straight line (or plane, or hyperplane, depending on the dimensionality) ☐
- ☐ Is a linear function of the parameters, but not necessarily of the input data because basis functions can be used to transform the data 1 pt. ☐
- ☐ None of the other options is correct

0 pt.

Vraag 8 **Beantwoord op:** 9 november 2021 - 09:05 **Duur:** 3 min. en 13 sec. **Score:** 1 van 1 pt. **1 pt.**

Given the following details of an articial neuron:

- $\mathbf{x}^T = [x_0, x_1, x_2]$, a vector collecting all inputs to the neuron, and in which x_0 is the bias.
- $\mathbf{w}^T = [w_0, w_1, w_2]$, a vector collecting all weights, and in which w_0 corresponds to the bias.
- y, a scalar value representing the true output for an input **x**
- f(z) = z, the activation function
- $o = f(\mathbf{x}^T\mathbf{w})$, the output value of the neuron
- $L(\mathbf{x},y,\mathbf{w}) = \sum_i (o^i - y^i)^2$ the loss function, where i iterates over all data points

If we consider just one data point ($\mathbf{x}^T = [1,2,4]$, $y = 3$), the weights set to $\mathbf{w}^T = [2,3,-1]$, and a learning rate of 0.01, which is the new value of w_2 if we perform one weight update using the gradient descent update rule?

- ☐ 0.08
- ☐ ☐ -1.08 1 pt. ☐
- ☐ -1.16
- ☐ -0.92

1 pt.

Vraag 9 **Beantwoord op:** 9 november 2021 - 09:10 **Duur:** 5 min. en 18 sec. **Score:** 1 van 1 pt. **1 pt.**

Antwoordmodel aangepast door G. Englebienne op 10 nov. 2021 11:08:32 met reden "The wrong answer was selected"

Consider a 6 dimensional dataset on which one applies PCA. The covariance matrix corresponding to the PCA's has the following elements on the diagonal 75.2; 11.0; 8.1; 6.2; 5.3; 4.2. How much variance is explained by the first two PCAs?

- 86.2%
- ✓ 73.8% 1 pt.
- 73.1%
- 60.3%

1 pt.

Vraag 10 **Beantwoord op:** 9 november 2021 - 10:13 **Duur:** 4 min. en 40 sec. **Score:** 1 van 1 pt. **1 pt.**

Artificial Neural Networks can be used in:

- Unsupervised Learning
- ✓ All answers are correct 1 pt.
- Supervised Learning
- Reinforcement Learning

1 pt.

Vraag 11 **Beantwoord op:** 9 november 2021 - 09:14 **Duur:** 29 sec. **Score:** 0 van 1 pt. **0 pt.**

Between autoencoders and PCA the following relation is true.

- All statements are true. 1 pt.
- If the autoencoder's activation functions are linear, it is very similar to PCA method.
- They are both feature representation learning methods.
- ✗ PCA is only linear transformation to the subspace while autoencoder is nonlinear transformation to the hidden units.

0 pt.

Vraag 12 **Beantwoord op:** 9 november 2021 - 09:17 **Duur:** 2 min. en 13 sec. **Score:** 1 van 1 pt. **1 pt.**

An articial neuron has three inputs x1, x2, and x3. Each input is connected to the neuron with a weight, w1, w2 and w3, respectively. Also, the neuron has a bias b and a sigmoid ($f(z) = \frac{1}{1+e^{-z}}$) activation function. Assuming that x1 = 1, x2 = 2, x3 = 0.5, w1 = 0.5, w2 = -1, w3 = -2, and b = 1 which is the neuron output value rounded to three decimals?

- 0.076
- 0.5
- 0
- ✓ 0.182 1 pt.

1 pt.

Vraag 13 **Beantwoord op:** 9 november 2021 - 10:16 **Duur:** 8 min. en 51 sec. **Score:** 0 van 1 pt. **0 pt.**

When choosing one feature from X1, ..., Xn while building a Decision Tree, which of the following criteria is the most appropriate to maximize? (Here, H() means an entropy, and P() means a Probability)

- ✗ P(Y|Xj)
- P(Y) - P(Y|Xj)
- H(Y|Xj)
- H(Y) - H(Y|Xj) 1 pt.

0 pt.

Vraag 14 **Beantwoord op:** 9 november 2021 - 09:27 **Duur:** 6 min. en 50 sec. **Score:** 0 van 1 pt. **0 pt.**

Consider a 1-dimensional dataset $A = \{5.0, 6.0, 7.0, 8.0, 9.0, 9.5, 9.6, 9.8\}$ and one is using for the conditional class likelihood $p(x|A)$ a Kernel density estimator of the form

$$p(x|A) = \frac{1}{N}\sum_{n=1}^{N} k(x - a_n) \text{ with } k(y) = 1 \text{ if } |y| \leq 1/2 \text{ and } k(y) = 0 \; otherwise$$

In the above expression N is the number of elements in A and a_n is the n-th element in A. Given the above what is the value for $p(9.2|A)$?

- 4/8
- ✓ 3/8 1 pt.
- ✗ 2/8

0 pt.

☐ 1/8

Vraag 15 **Beantwoord op:** 9 november 2021 - 09:36 **Duur:** 3 min. en 59 sec. **Score:** 0 van 1 pt. **0 pt.**

Suppose we want to maximize the function f(x,y)= $-3x^2 + 40x + 8xy + 288y - 10y^2$ subject to the constraint x + 2y = 58. What are the coordinates for the minimum? **0 pt.**

- ☐ (x,y) = (22,18) 1 pt.
- ☐ ☐ (x,y) = (2,28)
- ☐ (x,y) = (20,36)
- ☐ (x,y) = (28,15)

Vraag 16 **Beantwoord op:** 9 november 2021 - 09:36 **Duur:** 26 sec. **Score:** 1 van 1 pt. **1 pt.**

In Bayesian learning, we compute the following probability of a prediction y given an observation x and a training set $\{x, t\}$ using a set of model parameters : **1 pt.**

- ☐ $p(y|x, \hat{\theta})$ where we optimised $\widehat{\theta} = \text{argmax}_\theta \; p(\{x, t\}|\theta)$
- ☐ ☐ $p(y|x, \{x, t\})$, where the parameters are explicitly marginalised out 1 pt.
- ☐ $p(y|x, \hat{\theta})$ where we optimised $\widehat{\theta} = \text{argmax}_\theta \; p(\theta|\{x, t\})$
- ☐ None of the above

Vraag 17 **Beantwoord op:** 9 november 2021 - 10:21 **Duur:** 6 min. en 45 sec. **Score:** 0 van 1 pt. **0 pt.**

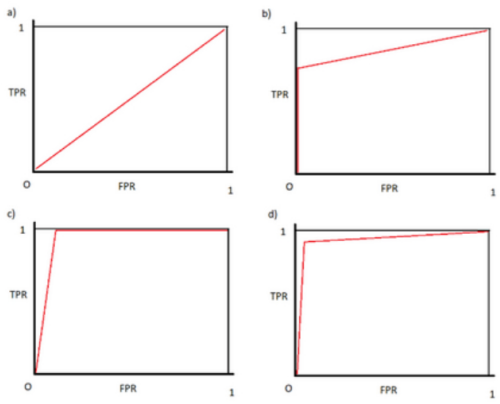
Let's consider a multilayer perceptron for classification with three hidden layers of neurons, ReLU activation function for the hidden neurons, and Mean Squared Error as loss function. The model is trained using standard stochastic gradient descent without momentum. If we initialize all weights and biases with zero what happens during training?

Choose the correct answer. **0 pt.**

- ☐ The model will not be capable to learn. 1 pt.
- ☐ ☐ The model will learn to classify correctly the input data if enough training time is given.
- ☐ The given details are not sufficient to give an answer.
- ☐ The model will overfit on the training data.

Vraag 18 **Beantwoord op:** 9 november 2021 - 09:43 **Duur:** 7 min. en 58 sec. **Score:** 0 van 1 pt. **0 pt.**

Assume that you are in situation were you have to develop and deploy a Machine Learning model for detecting a deadly disease. One of the requirments for the model is that it may not miss an infected person because the disease is very infectious. You trained for models A, B, C and D with the following ROC curves.



Given the above requirement, which model performs best? **0 pt.**

- ☐ Model A
- ☐ Model B
- ☐ Model C 1 pt.
- ☐ ☐ Model D

Vraag 19 **Beantwoord op:** 9 november 2021 - 09:44 **Duur:** 18 sec. **Score:** 1 van 1 pt. **1 pt.**

Which statement about K-fold cross validation is true?

1 pt.

- Every data element is used 1 time for testing and K times for training.
- ☑ Every data element is used 1 time for testing and K-1 times for training. 1 pt.
- Every data element is used K-1 times for testing and 1 time for training.
- Every data element is used K times for testing and K times for training.

Vraag 20 **Beantwoord op:** 9 november 2021 - 10:24 **Duur:** 6 min. en 32 sec. **Score:** 1 van 1 pt.

1 pt.

Assume that:
1: We have a two class classication problem in a 4-dimensional space.
2: We apply Bayes law to estimate $P(C_k|x), k = 1, 2$.
3: We assume that the likelihoods are modelled by normal (Gaussian) probability distributions with shared diagonal covariance matrices.

How many parameters does one need to estimate or learn from the data?

1 pt.

- 14
- ☑ 13 1 pt.
- 9
- 19

Vraag 21 **Beantwoord op:** 9 november 2021 - 09:52 **Duur:** 2 min. en 4 sec. **Score:** 1 van 1 pt.

1 pt.

Given the regression model $h_\theta(x)$ and the cost function $J(\theta)$, as defined next:

$h_\theta(x) = \theta_3 x^3 + \theta_2 x^2 + \theta_1 x + \theta_0$, where $\theta_3, \theta_2, \theta_1, \theta_0$ are the model parameters

$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^2$, where i iterates over all data, $x^{(i)}$ represents the input of a data point (i), and $y^{(i)}$ represents the true output of a data point (i)

Which is the partial derivative of the cost function with respect to variable θ_3, i.e., $\frac{\partial J(\theta)}{\partial \theta_3}$?

1 pt.

- ☑ $\frac{\partial J(\theta)}{\partial \theta_3} = \frac{2}{n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)^3}$ 1 pt.
- $\frac{\partial J(\theta)}{\partial \theta_3} = \frac{1}{2n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})^3 x^{(i)^3}$
- $\frac{\partial J(\theta)}{\partial \theta_3} = \frac{1}{n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$
- $\frac{\partial J(\theta)}{\partial \theta_3} = \frac{1}{n}\sum_{i=1}^{n}(h_\theta(x^{(i)}) - y^{(i)})x^{(i)^3}$

Vraag 22 **Beantwoord op:** 9 november 2021 - 10:32 **Duur:** 10 min. en 51 sec. **Score:** 0 van 1 pt.

0 pt.

Which of the following statements is true:

0 pt.

- Training a probabilistic model by maximum likelihood (ML) parameter estimation requires iterative gradient descent and is likely to cause overfitting.
- Training a probabilistic model by maximum a posteriori (MAP) parameter estimation with a proper prior is much more computationally complex than ML estimation, but solves the problem of overfitting.
- ☑ Training a probabilistic model by Bayesian learning (marginalising out the parameter values) is both expensive and leads to overfitting, but is elegant and formally the correct thing to do.
- None of the above is true. 1 pt.

Vraag 23 **Beantwoord op:** 9 november 2021 - 09:57 **Duur:** 1 min. en 18 sec. **Score:** 1 van 1 pt.

1 pt.

☑ Antwoordmodel aangepast door G. Englebienne op 10 nov. 2021 11:15:38 met reden "It was not clear what information gain to use, so we accept both the correct answer (with GINI) and "none of the above" if entropy was used."

Given the following table of observations, calculate the information gain IG(X|Y) that would result from learning the value of X.

1 pt.

X	Y
Red	$True$
$Green$	$False$
$Blue$	$False$

$Blue \quad True$

- ☐ 1/2 — 1 pt.
- ☐ 1
- ☐ None of them — 1 pt.
- ☐ 3/2

Vraag 24 **Beantwoord op:** 9 november 2021 - 09:59 **Duur:** 2 min. en 13 sec. **Score:** 0 van 1 pt. — 0 pt.

The *chain rule* is a key element in:

Choose the correct answer.

- ☐ all answers are correct — 1 pt.
- ☐ convolutional neural networks training
- ☐ multilayer perceptron training
- ☐ backprogation

0 pt.

Vraag 25 **Beantwoord op:** 9 november 2021 - 09:59 **Duur:** 1 min. en 23 sec. **Score:** 0 van 1 pt. — 0 pt.

Deep Learning algorithms perform representation learning and can be used for high dimensional data as an alternative to:

- ☐ manual feature engineering — 1 pt.
- ☐ Stochastic Gradient Descent
- ☐ an artificial neuron
- ☐ ReLU activation function

0 pt.

Vraag 26 **Beantwoord op:** 9 november 2021 - 10:40 **Duur:** 10 min. en 5 sec. **Score:** 1 van 1 pt. — 1 pt.

What is the purpose of regularization?

- ☐ Maximising the performance of the model on unseen data to reduce overfitting.
- ☐ Training the model on k different folds of the data to reduce overtting.
- ☐ Reducing the number of weights to reduce overfitting.
- ☐ Preventing large weight values to reduce overfitting. — 1 pt.

1 pt.

Vraag 27 **Beantwoord op:** 9 november 2021 - 10:04 **Duur:** 41 sec. **Score:** 0 van 1 pt. — 0 pt.

Which statement is false?

- ☐ The SVM algorithm computes the maximum margin hyperplane.
- ☐ The kernel trick allows us to use support vector machines as a loss function in neural networks. — 1 pt.
- ☐ By using the kernel trick, the SVM algorithm can efficiently perform non-linear classification.
- ☐ The kernel trick allows us to compute SVMs in a high dimensional space.

0 pt.

Vraag 28 **Beantwoord op:** 9 november 2021 - 10:05 **Duur:** 2 min. en 51 sec. **Score:** 1 van 1 pt. — 1 pt.

When using stochastic gradient descent to optimize the parameters of a linear regression or a multilayer perceptron model, what happens if we use a fixed learning rate which is too large?

Choose the correct answer.

- ☐ The model will converge in normal time.
- ☐ The model will converge very slowly.
- ☐ There are no problems, as gradient descent is robust to the choice of the learning rate.
- ☐ The model may fail to converge. — 1 pt.

1 pt.