Exam B

I. True or False(45 statements = 9.0v)

Please mark the following statements as **T**rue or **F**alse (+0.2 correct, –0.1 wrong):

Group I. Classification

T 1. A 1-NN classifier has always zero training error. F

T 2. Increasing the depth of a decision tree cannot increase its training error. F

F 3. A negative observation that is wrongly labelled is termed false negative. FP

T 4. In theory, a decision tree learned from data with m binary attributes can represent any Boolean function over those m attributes if enough observations are provided. V

T 5. In cross-validation, the higher the number of folds, the higher the number of training observations per fold.

Group II. Clustering

F 6. Hamming distance is adequate to handle ordinal data with high cardinality. categorical

F 7. A rand index close to zero suggests that the clustering algorithm was not able to guarantee high cluster dissimilarity.

T 8. Purity is biased when the number of found clusters approaches the total number of observations. V

T 9. Both k-median and k-medoids are more robust to outliers than k-means. V

T 10. Complete link criterion tends to break large clusters and is biased towards globular clusters.

Group III. Pattern mining

T 11. The monotonic property states that the subsets of a frequent itemset are also frequent. → too, entropy C, upper part entropy C

F 12. Considering the use of equal-frequency/depth discretization, the prior normalization of an attribute affects the produced bins.

T 13. The assessment of lift is particularly relevant since the rule's consequent can appear on transactions without the rule's antecedent. PE $lift(A \to B) = P(B|A) / P(B)$

F 14. A closed itemset is always a maximal itemset.

T 15. Assuming the lengthiest pattern is p, then Apriori performs at most $O(p)$ database scans.

Group IV. Biclustering

Consider the following dataset and biclusters:

	y_1	y_2	y_3	y_4
x_1	1	1	0	2
x_2	1	2	1	3
x_3	1	2	1	3
x_4	4	1	1	0

γ_B	0	1	1
	y_2	y_3	y_4
x_1	1	0	2
x_2	2	1	3
x_3	2	1	3

$\varphi_B = 0$, $\varphi_B = 1$, $\varphi_B = 1$

T 16. $B=(I=\{x_2,x_3\},J=\{y_2,y_3,y_4\})$ is an order-preserving bicluster.

T 17. The largest constant bicluster with $I=\{x_2,x_3\}$ has pattern of length $|\varphi_B|=4$. ?

T 18. $B=(I=\{x_1,x_2,x_3\},J=\{y_2,y_3,y_4\})$ is an additive bicluster with shifting factors $\gamma_B=\{\gamma_1=0,\gamma_2=1,\gamma_3=1\}$.

T 19. $B_1=(I_1=\{x_1,x_2,x_3\},J_1=\{y_1,y_2\})$ and $B_2=(I_2=\{x_2,x_3,x_4\},J_2=\{y_2,y_3\})$ are plaid biclusters with pattern $\varphi_{B_1}=\{c_1=1,c_2=1\}$ and $\varphi_{B_2}=\{c_1=1,c_2=1\}$. ?

F 20. Constant bicluster $B=(I=\{x_1,x_2,x_3\},J=\{y_1,y_2,y_3\})$ with $\eta_{ij} = 0$ has a quality of 0.75.

	y_1	y_2
x_1	1	1
x_2	1	2
x_3	1	2

	y_2	y_3
x_2	2	1
x_3	2	1
x_4	1	1

$\frac{9-2}{9} = \frac{7}{9}$

Group V. Data reduction

F 21. Filter procedures for feature selection are commonly applied with accuracy measures.

T 22. Learning curves can be considered to estimate the best number of features to select.

T 23. Spearman correlation is preferred over Pearson correlation if the order of quantities is more relevant than their absolute value.

F 24. Generally, backward subset selection is more computationally expensive than forward subset selection if the majority of features are non-redundant and discriminative.

F 25. Linear discriminant analysis (LDA) finds the axes that show greatest variation for the observations of each class.

Group VI. Time series data

T 26. When applying the discrete Fourier transform (DFT) to analyze a signal with a fixed sampling rate, the time window (number of time points) needs to be high enough to model low frequencies.

T 27. A short discrete Fourier transform (SDFT) is a DFT on sliding segments of a signal.

F 28. SDFT is preferred over classic DFT when a time series is stationary.

T 29. While wavelets, such as Haar wavelets, are more appropriate to understand household electricity consumption signals, sinusoids are more appropriate to understand brain signals.

F 30. Codebook representations of time series are symbolic representations that produce time series with higher dimensionality than SAX representations.

Group VII. Regression

F 31. Given attributes y1, y2 and y3, if covariance (in absolute value) between y1 and y2 is higher than covariance between y1 and y3, then y1 and y2 have higher correlation.

T 32. Given a linear regression model, if a scatter plot shows that the residuals are highly correlated (Pearson correlation close to 1), then the learned model is not good.

T 33. Multiple linear regression is sensitive to outliers.

T 34. Decision tree regressors can only estimate as many quantities as the number of leaves.

F 35. AUC can be also considered to evaluate regression models.

Group VIII. Mining complex data

T 36. Sequential pattern mining can be applied both on symbolic time series data and itemset sequence data

F 37. The orders between the items in a sequential pattern can be partially defined

T 38. Given a dataset with n time series, to learn a tabular data encoding using regression coefficients: there is the need to learn as many regressions as the number of observations

F 39. The spatial slicing principle for spatiotemporal data analysis is verified when spatial content can be separated from the remaining static and temporal content during the learning

F 40. Task partitioning procedures for distributed data mining aim to partition data into subsets and distribute their analysis across processors

Group IX. Pre-processing

F 41. Outlier analysis aims to detect attributes with values deviating significantly from expectations.

T 42. In semi-supervised outlier analysis, it is more relevant to know non-outliers than outliers.

T 43. Statistical approaches to outlier detection assume data to be generated by a distribution to test outlier likelihood.

T 44. Clustering is an effective means to perform subsampling when $n_{new}/n_{original} \ll 1$ (where n_{new} and $n_{original}$ are respectively the number of observations in the final and original dataset).

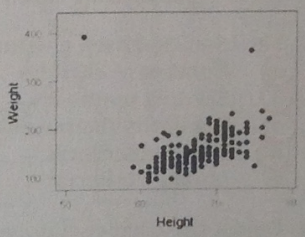
T 45. Binning numeric variables and merging categoric values is a way of reducing domain cardinality.

II. Multiple Choice (12 questions 0.4v each = 4.8v)

Select **all the true answers** (none, one or more than one answer are allowed). The grade corresponds to k/n, with n the number of correct answers and k the number of correct options taken. Wrong answers discount 1/2n.

Group I. Regression

1. The difficulties to learn a regression model to analyze the following dataset are:



   a. Presence of one or more outliers
   b. Differences in scale between variables
   c. Non-normalized data: height values higher than 50
   d. Curvilinear data
   e. Response variable is not quantitative

2. Which of the following criteria contribute to a smoothed regression model?
   a. Increasing depth of decision trees
   b. Increasing k of nearest neighbours
   c. Parameterizing nearest neighbours with uniform weights instead of distance-based weights
   d. Application of linear and kernel smoothers

3. Identify true statements on single-wise (versus local/piece-wise) regression models:
   a. *k*NN is a single-wise regressor
   b. Decision tree is a single-wise regressors
   c. Single-wise regressors typically have a higher generalization ability (low complexity term)
   d. Single-wise regressors typically suffer from underfitting risks

4. Why is PCA sometimes used as a preprocessing step before regression?
   a. To select predictors
   b. To minimize overfitting risks
   c. To expose information missing from the input data
   d. To make computation faster by reducing the dimensionality of the data

Group II. Classification

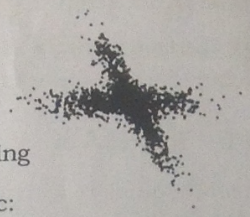
5. Consider the problem of building decision trees with *k*-ary splits (split one node into *k* nodes) with entropy impurity. Which of the following is/are true?
   a. The algorithm will always choose $k = 2$
   b. The algorithm will prefer high values of *k*
   c. There will be *k*-1 thresholds for a *k*-ary split
   d. This model is strictly more powerful than a binary decision tree

6. Which of the following are true about each individual tree in a random forest?
   a. Individual tree is built on a subset of the features
   b. Individual tree is built on all the features
   c. Individual tree is built on a subset of observations
   d. Individual tree is built on full set of observations

7. Why would we use a random forest instead of a decision tree?
   a. For lower training error
   b. To reduce underfitting propensity
   c. To reduce overfitting propensity
   d. To better approximate posterior probabilities
   e. To facilitate human interpretability

8. Consider the learning of a classifier from a dataset with 1000 attributes. 50 of them are discriminative. Another 50 features are direct copies of the first 50 attributes. The final 900 features are not informative. Assume there is enough data to reliably assess how useful features are:
   a. 100 features will be selected by mutual information filtering
   b. 100 features will be selected by a backward wrapper method
   c. 50 features will be selected by mutual information filtering
   d. 50 features will be selected by a forward wrapper method

Group III. Clustering



9. Given the following data, the true clusters are better described by:
- **a.** Model-based clustering than density-based clustering
- **b.** Soft clustering than deterministic clustering
- **c.** Classic partition-based clustering than fuzzy clustering
- **d.** Agglomerative-based clustering (single link) than model-based clustering

10. Partitioning-based clustering algorithms can be parameterized with specific:
- **a.** Number of clusters
- **b.** Seeding methods
- **c.** Similarity distances
- **d.** Centroid criteria
- **e.** Linkage criteria

$\Sigma(x_i - \hat{x})^2$

11. Select the correct statements on the sum of squared errors (SSE):
- **a.** SSE is a measure of clustering separation
- **b.** SSE is not adequate to compare clustering solutions with a different number of clusters
- **c.** If SSE on a target dataset is ρ and the SSE expectations on randomized data is always lower than ρ, then the clustering solution is statistically significant regarding SSE
- **d.** If SSE on a target dataset is ρ and the SSE expectations follow a Gaussian distribution X and P(X<ρ)=1E-3, then the clustering solution is statistically significant regarding SSE

12. Select the advantages of clustering with minimum linkage (in contrast with maximum linkage)
- **a.** Ability to identify large clusters or different sizes
- **b.** Ability to identify clusters with non-elliptical shapes
- **c.** Robustness to outliers and noise
- **d.** Ability to model overlapping clusters

III. Calculus (14 questions = 6.2v)

Group I. Time Series [2.4v]

Considering the following time series data
(where time series are assumed to be already normalized and DTW is applied with squared loss):

	time series	**z**
x_1	<1, -1, 0>	5
x_2	<0, -2>	6
x_3	<-1, 1, 1>	7
x_4	<-1, 1, 0, 2, 3>	3

5b) $RMSE = \sqrt{\frac{1}{n}\Sigma(z-\hat{z})^2} = \sqrt{\frac{1}{5}(4+1+1+4)} = \sqrt{\frac{10}{5}} = \sqrt{2}$

1. [0.2v] What is the PAA representation for x_2 with 3 segments? <0, -1, -1> X
2. [0.5v] What is the DTW cost between x_1 and x_3? How many alignments has this path? ✓
3. [0.6v] Assuming a KNN with k=2 trained over $\{x_2, x_3, x_4\}$, and distances $d(x_1, x_2)=2$, $d(x_1, x_3)=3$, $d(x_1, x_4)=1$. What is the estimated quantity for x_1 when considering k=2 and distance weights? mais perto: x_2 e x_4 → $\hat{z}_1 = (6+3)/2 = 9/2 = 4{,}5$ X
4. [0.4v] Considering time series x_4 observed between t=1 and t=5. Assuming one differencing operation, what is the naïve forecast for t=7?
5. [0.7v] Considering time series x_4 observed between t=1 and t=5. Assuming a time series regression with $\hat{\beta}$=[$\hat{\beta}_0$=2, $\hat{\beta}_1$=1]: identify a) the residuals on the observed series, and b) the associated RMSE.

2.

	1	4	8	(6)
	1	4	8	[5]
x3	-1	[4]	[4]	5
		1	-1	0
		x1		

DTW cost = 6
4 Alignments

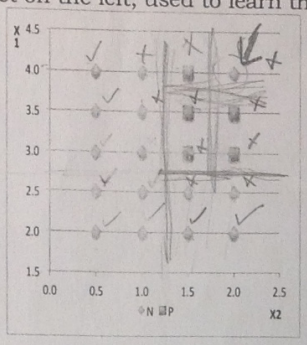
4. <-1, 1, 0, 2, 3>
<2, -1, 2, 1>
$\hat{x}(5) = 3$
$\hat{x}(6) = 3+1 = 4$
$\hat{x}(7) = 4+1 = 5$ ✓

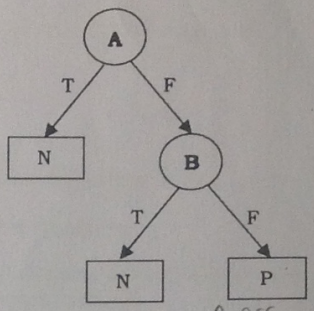
5. $\hat{z} = \beta_0 + \beta_1 x = 2 + x$
$\hat{z}_0 = 2-1 = 1$ $\hat{z}_2 = 2+0 = 2$
$\hat{z}_1 = 2+1 = 3$ $\hat{z}_3 = 2+2 = 4$
$\hat{z}_4 = 2+3 = 5$

a) residuals:
<3-1, 3-3, 3-2, 3-4, 3-5>
<2, 0, 1, -1, -2>

Group II. Classification [2.5v]

Given the dataset on the left, used to learn the tree on the right with a pre-pruning strategy:





6. [0.5v] What are the tests performed in A and B when using information gain criteria to learn the tree?
7. [0.5v] What is the tree accuracy? And sensitivity for the negative class (N)?
8. [0.5v] How many instances is KNN (with k=3) able to correctly classify in the dataset, using the leave-one-out strategy?
9. [0.5v] Consider a random forest learnt using C4.5 to train decision stumps on the dataset. How many different classifiers would be trained?
10. [0.5v] How would that random forest classify the instance (4.0, 2.0)?

Group III. Data Reduction [1.3v]

Consider the following eigenvalues and eigenvectors (ignore the fact $\|v_i\|=1$) produced from a dataset with 3 attributes:

$\lambda_1=2.5$
$\lambda_3=1$
$\boldsymbol{v}_1=[\ 2\ \ 1\ -2\ ]$
$\boldsymbol{v}_2=[\ 1\ -2\ \ 0\ ]$

and a covariance matrix given by $\begin{bmatrix}4 & 1 & 2\\ 1 & 2.5 & 1\\ 2 & 1 & 5\end{bmatrix}$

11. [0.3v] What is the value for eigenvalue λ_2?
12. [0.3v] Assuming $\lambda_2=1.5$, what is the explained variability by the two first components?
13. [0.3v] Considering a (centred) observation with values x_1 = [1 1 1] and the application of PCA with the two first components, what are the component values for x_1?
14. [0.4v] In the same conditions, what are the recovered values of the first observation using the inverse PCA with two components?

13. [illegible] = Data transformed = FeatureVector^T x DataAdjusted

14. recovered values = (Transformed Data x Feature vector) + original mean

$= \begin{bmatrix}4 & 1 & 2\end{bmatrix}$

I. Calculus (14 questions = 8.85v)

Group I. Regression and Clustering [3.15v]

Consider the dataset below and answer the questions:

	y1	y2	*cluster*	*class*	*z*
x1	1	2	C1	A	5.5
x2	3	4	C1	A	9.0
x3	0	1	C1	B	5.0
x4	1	0	C2	B	6.5

x1 x2 x3 (x4)

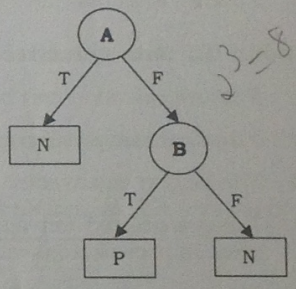
1. [0.8v] Given the clusters C1 and C2 in the dataset, what is the silhouette index for x1 using Chebyshev distance? And Manhattan distance? $q=1$ 0.25 ; 1/3
   ↳ max ↳ 1 1
2. [0.85v] Given the ground truth (*class*), how much is the purity of the clustering solution (*cluster*)? And the rand index? 3/4 ; 1/2
3. [0.5v] Considering uniquely attribute y1, separate observations in two clusters using agglomerative clustering with maximum link. x1 x3 x1 x4
4. [0.6v] What is the *residual* value associated with observation x1 assuming amultiple linear regression with $\hat{\boldsymbol{\beta}}=[\hat{\beta}_0=3, \hat{\beta}_1=1, \hat{\beta}_2=2]$. −2,5
5. [0.4v] Assuming a constant regression model $\hat{z}=6$, calculate the MAE. 5/4

Group II. Classification [3.6v]

16 instances

Consider the following dataset (note that the shadowed cells are contradictory among them).

A	F	F	F	F	F	F	F	F	T	T	T	T	T	T	T	T
B	F	F	F	F	T	T	T	T	F	F	F	F	T	T	T	T
C	F	F	T	T	F	F	T	T	F	F	T	T	F	F	T	T
Class	N	N	N	N	N	N	N	N	N	N	N	P	P	N	P	P



1. [0.5v] With which of the studied algorithm(s) is it possible to learn the decision tree in the figure? TODOS
2. [1.0v] Consider a random forest trainedon the given dataset and using C4.5 to train regular decision trees (trees with three nodes: the root and two siblings testing the same attribute). How many different classifiers would be trained? 2
3. [1.2v] How does naïve Bayes classify the instance (A=T,B=T,C=F)? Present the values for P(A=T|P) and P(A=T|N). P ; 1 / 1/3
4. [0.5v] Does 1-NN classify the same instance correctly using the leave-one-out strategy? P N NO
5. [0.4v] And without removing the instance, is it possible to classify it? NO

Group III. Pattern Mining [2.1v]

Given a transaction dataset, where the only patterns are (ABDE) s=5%, (BCE) s=7% and (ACDE) s=9%.

1. [0.3v] How many times does Apriori algorithm scans the dataset? 4
2. [0.7v]For which 3-candidates, the Apriori algorithm counts the support for...

D. All generated candidates	E. Only for the proper maximal subsets of each 4-pattern	F. It's not possible to answer

3. [0.8v] Consider that the three patterns are frequent itemsets found from a set of sequences, with the corresponding supports. What is the possible maximum support for the sequence **(AB)(BCE)D**? 5%
4. [0.3v] In the same conditions, can **(CE)CE** be a frequent sequence? yes

II. True or False(40 statements = 8v)

Please mark the following statements as **T**rue or **F**alse (+0.2 correct, –0.1 wrong):

Group I. Pattern mining and Biclustering

F 1. The anti-monotonic property states that supersets of a frequent itemset are infrequent.

F 2. The lift measure of an association rule A⇒B does not change if we add a new transaction that does not contain either A or B. $lift(A \to B) = P(B|A) / P(B)$

3. The pattern AB(A,B)AC cannot be discovered by PrefixSpan.

F 4. Given a dataset and a bicluster in it, a false positive bicluster is a statistically significant one that was not found.

F 5. A biclustering solution with 2 biclusters with overlapping cells is always non-exhaustive on rows and columns.

Group II. Classification

T 1. Given an unbalanced dataset, a classifier with 90% testing accuracy is always more useful than a classifier with 80% testing accuracy.

T 2. Naive Bayes is a linear classifier (linear boundaries to separate observations).

T 3. All attributes are equallyimportant in Naïve Bayes.

F 4. Normalizing attributes improve the performance of any classifier.

T 5. Feature selection does not usually improve the performance of KNN.

Group III. Data reduction

T 1. Feature selection can be applied supervisedly with a numeric output variable.

2. Principal component analysis is a centered singular value decomposition.

F 3. The largest eigenvector of the covariance matrix is the direction of minimum variance in the data.

T 4. The generalized forward subset selection greedily adds a fixed number of features per iteration that most improves cross-validation accuracy.

T 5. Given a *m*-dimensional dataset, PCA can reconstruct any data point using m-1 components of PCA with zero reconstruction error.

Group IV. Clustering

1. The sum of diagonal entries in a pairwise distance matrix equals one.

F 2. When pairs of observations are known to belong to the same cluster, we face a semi-supervised clustering task.

F 3. In k-medoids, the centroid is given by the mode of categorical attributes and median of numerical attributes.

T 4. Agglomerative clustering algorithms allow to decide the number of clusters after clustering is done.

F 5. k-means does not adequately identify spherical groups of observations.

Group V. Regression

T 1. The higher the covariance (in absolute value) between two attributes, the higher their correlation.

T 2. A linear regression with more than one dependent variable is called multiple linear regression.

T 3. Cross-validation can be applied to minimize the overfitting propensity of a regression model.

T 4. A logistic regression model is a classification model.

T 5. By minimizing regression coefficients, Lasso estimation is useful to discard attributes that do not help to estimate the output variable.

Group VI. Pre-processing

T 1. An outlier can be either inconsistent with the remaining data or with its neighbourhood.

T 2. Clustering can be applied to perform outlier analysis and subsampling.

3. When assessing a classifier, the imputation of missing values should be always performed prior to cross-fold validation as long as imputation does not depend on the class.

4. In proximity-based approaches for outlier analysis, either outliers have distant nearest neighbours or density around outliers differs from density around neighbours.

T 5. Normalizing attributes **can** affect the outcome of an equal-width/range discretization.

Group VII. Time series data

T 1. When applying the DFT, the sampling rate at which a signal is measured needs to be high enough to adequately model low frequencies.

T 2. While DWT (wavelet) offers a temporal-based decomposition of a signal, a short DFT (Fourier) (SDFT) strictly offers a frequency-based decomposition.

3. Contrasting with SAX, codebooks encode motifs discovered using multiple temporal-resolutions.

T 4. In addition to seasonal variation, time series can be described by a cyclical variation component.

T 5. A time series has a linear trend if the *p*-value of Dickey-Fuller test is low (typically less than 0.01).

Group VIII. Complex data mining

F 1. Minkowski distances are more adequate than DTW for time series clustering if we do not want to tolerate temporal misalignments.

T 2. Associative classifiers for spatiotemporal data rely on spatiotemporal pattern mining.

T 3. Chords and phrases are temporal patterns for univariate time series data.

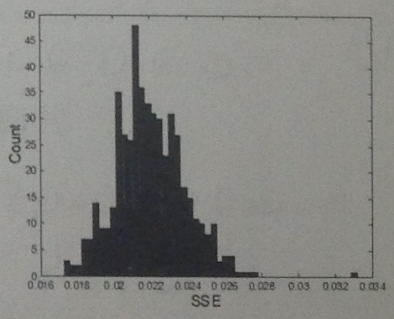
T 4. Horizontal data partitioning principles can be used to distribute the learning of decision trees since the information gain of each data attribute is independently tested.

F 5. *k*NN is always able to efficiently and incrementally learn from data streams as long as *k* is less than the number of simultaneously arriving observations.

III. Multiple Choice (9 questions 0.35 each = 3.15v)

Select the **only true answer** (just one). Wrong answers discount half the grade.

Group I. Clustering



1. Consider the following analysis of sum squared errors gathered from a thousand of randomized datasets using *k*-means:
   a. A SSE in [0.02,0.023] is statistically significant
   b. A SSE above 0.34 is statistically significant
   c. A SSE below 0.017 is statistically significant
   d. None of above

2. Which of the following is **not** applicable to the *k*-Means algorithm:
   a. Dependent on good initialization/seeding ✓
   b. Sensitive to outliers and noisy data ✓
   c. Not suitable to discover clusters with non-convex shapes ✓
   d. Not able to separate clusters when their variance is small in all directions

3. When we are interested in generating overlapping cluster membership, we should use:
   a. *k*-medoids clustering ✗
   b. Hierarchical clustering
   c. Model-based clustering
   d. Density-based clustering

Group II. Classification

1. In which of the following scenario a gain ratio is preferred over Information Gain?
   a. When a categorical variable has high cardinality
   b. When a categorical variable has low cardinality
   c. Numeric variables
   d. None of above

2. A random forest with low training error is getting abnormally bad performance on the validation set. What could be causing the problem?
   a. Decision trees are too weak
   b. Randomly sampling too few features when choosing a split
   c. Too few trees in the ensemble
   d. None of above

   ~~random forest~~ ensembles

3. What does it mean to perform a data bootstrap?
   a. To sample m features with replacement from the total m
   b. To sample m features without replacement from the total m
   c. To sample n examples with replacement from the total n
   d. To sample n examples without replacement from the total n

Group III. Others

1. Which of the following factors does **not** contribute to increase the average size of biclusters:
   a. Increasing tolerance to noise
   b. Given perfect quality, increasing the cardinality of attributes in discrete data
   c. Looser coherence strength (higher deviations allowed) in real-valued data
   d. More flexible coherence assumptions (e.g. choosing additive instead of constant assumption)

2. Which of the following is **not** a typical property of smoothed regression models (such as a multiple linear regression model):
   a. Small training error
   b. Low overfitting risk
   c. Low complexity
   d. Interpretability

3. Given the already normalized time series x1=<-2,0,2> and x2=<2,1>, select the correct answer:
   a. PAA representation of x1 with length 2 is <-1,1>
   b. SAX representation of x2 using two symbols is <b, a>
   c. DTW with Manhattan loss between x1 and x2 is 6
   d. The number of DTW alignments between x1 and x2 is 4

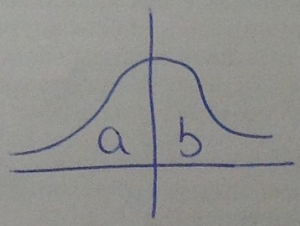
PAA <-2,0,2> 3 segments $\frac{3}{2} = 1,5$

↓ 2 segments

<-2x1+0x0,5 , 0x0,5+2x1>

<-2 , 2>

(a,b)



1	7	5	6
2	4	6	6
x2 / x1	-2	0	2

1-2-21