Past Paper

20_Main

Question 1

1. Describe two ways in which the tendency of decision trees to overfit their training data can be overcome?
   描述两种克服决策树过度拟合训练数据倾向的方法？
   - Restrict the tree depth
     限制树深度
   - Use in an ensemble, eg, through boostingm or in a random forest
     在集成中使用，例如通过 boostingm 或在随机森林中使用
2. Provide the mathematical definition of information entropy and sort the following binary strings from lowest to highest entropy?
   提供信息熵的数学定义并将以下二进制字符串从最低熵到最高排序？
   1. 1101101110
   2. 1101010010
   3. 1001000101
   - Information entropy $S = \sum_i p(i) \ln p(i)$.
   - The information entropy of the three strings is:
     - $p(0) = 0.3, p(1) = 0.7 \to S = -0.3 \ln 0.3 - 0.7 \ln 0.7 = 0.611$
     - $p(0) = 0.5, p(1) = 0.5 \to S = -0.5 \ln 0.5 - 0.5 \ln 0.5 = 0.693$
     - $p(0) = 0.6, p(1) = 0.4 \to S = -0.6 \ln 0.6 - 0.4 \ln 0.4 = 0.673$
       so the order is i-iii-ii
3. The following table describes a binary classification dataset $D = \{x_i, y_i, z_i, t_i\}_{i=0}^{7}$ with independent variables x, y, and z; and dependent variable t

i	xi	yi	y'i	ti
0	0	0	0	0
1	0	0	1	0
2	0	1	0	0
3	0	1	1	1
4	1	0	0	0
5	1	0	1	1
6	1	1	0	1
7	1	1	1	1

The data generating process returns $t = 1$ when two or more of the independent variables are 1
Using the principle of maximum information gain to determine the order of the variable splits, construct the full decision treee for this dataset, using data points $i = \{0, 1, 2, 4, 6, 7\}$ as the training set. Test your tree on data points $i = \{3, 5\}$ and comment on your result
You may find the following table of logarithms helpful.

x	1/4	1/3	1/2	2/3	3/4	1
ln(x)	-1.386	-1.099	-0.693	-0.405	-0.288	0

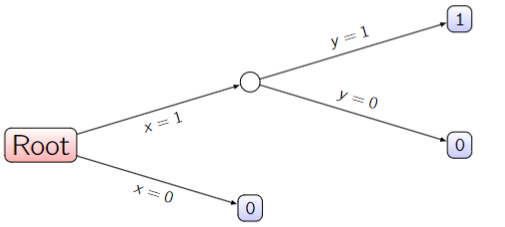
- First, we compute the parent entropy on the whole set, which has four zeros and two ones. The entropy of this is $S = -\frac{1}{3} \ln \frac{1}{3} - \frac{2}{3} \ln \frac{2}{3} = 0.637$
- Now we split on each of the variables:
  - x_0: three zeros gives three zeros; three ones give one zero and two ones. The entropy of this is $S = \frac{1}{2}(-1 \ln 1 - 0 \ln 0) + \frac{1}{2}(-\frac{1}{3} \ln \frac{1}{3} - \frac{2}{3} \ln \frac{2}{3}) = 0.318$
  - x_1: This is the same as x_0
  - x_2: four zeros gives three zeros and one one; two ones give ones give one zero and one ones. The entropy of this is $S = \frac{2}{3}(-\frac{3}{4} \ln \frac{3}{4} - \frac{1}{4} \ln \frac{1}{4}) + \frac{1}{3}(-\frac{1}{2} \ln \frac{1}{2} - \frac{1}{2} \ln \frac{1}{2}) = 0.51$
- So the first split in the data should be made on either x or y
  - Choice 1: split on x. Selecting the samples where $x = 0$ we have:

i	xi	yi	zi	ti
0	0	0	0	0
1	0	0	1	0
2	0	1	0	0

This branch is homogeneous with entropy $S = 0$ and so it cannot be split further
Selecting the samples where $x = 1$ we have:

i	xi	yi	zi	ti
4	1	0	0	0
6	1	1	0	1
7	1	1	1	1

The parent state entropy is $S = -\frac{1}{3} \ln \frac{1}{3} - \frac{2}{3} \ln \frac{2}{3} = 0.637$
Splitting on y: one zero gives one zero; two ones give two ones. The entropy $S = 0$
Splitting on z: two zeros gives one one, one zero; one one gives one one. The entropy is $\frac{2}{3}(-\frac{1}{2} \ln \frac{1}{2} - \frac{1}{2} \ln \frac{1}{2}) + \frac{1}{3}(-\ln 1) = 0.462$
Therefore, we split on y. This gets us to homogeneous leaf nodes and there is no need to split on z. The resulting tree is therefore:



Question 2

1. L_2 regularisation is sometimes used to control the solution to regression problems. It is often referred to as a shrinkage method.
   L_2 正则化有时用于控制回归问题的解决方案。 它通常被称为收缩方法。

    1. What is meant by the term "shrinkage method"?
       “收缩法”是什么意思？
        - Shrinkage methods encourage model weights to be kept small
          收缩方法鼓励模型权重保持较小
    2. Explain why L_2 regularisation is classified as a shrinkage methode, with reference to its probabilistic interpretation.
       参考其概率解释，解释为什么 L_2 正则化被归类为收缩方法。
        - It tries to minimise the sum of the squares of the model weights. This can be shown to be equivalent to imposing a Gaussian prior of mean zero on the weights and so large weights are unlikely
          它试图最小化模型权重的平方和。 这可以证明相当于在权重上施加均值为零的高斯先验，因此不太可能有大的权重
    3. Give another example of a shrinkage method and explain how it influences the solutions of regression problem
       给出收缩方法的另一个例子并解释它如何影响回归问题的解决方案
        - L_1 is one examples, this tend tends to encourage sparse weights(mostly zeros)
          L_1 就是一个例子，这往往会鼓励稀疏权重（大部分是零）

2. The expectation value of the least squares loss function can be written as

$$E[L] = \underbrace{\sigma^2}_{i} + \underbrace{var[f]}_{ii} + \underbrace{(h - E[f])^2}_{iii}$$

   Explain the meaning of the symbols $\sigma . f, and\ h$; and of the terms i, ii, and, iii

    - σ: Standard deviation of data generating process
      σ: 数据生成过程的标准差
    - f: estimated function
      f：估计函数
    - h: Mean of the data-generating process
      h：数据生成过程的平均值
    - Term i: implicit loss due to measurement uncertainty
      第 i 项：由于测量不确定性造成的隐含损失
    - Term ii: the variance in the estimated function as a consequece of the measurement uncertainty
      第 ii 项：由于测量不确定性而导致的估计函数的方差
    - Term iii: the difference between the estimated function and the mean of the true data generating process
      第三项：估计函数与真实数据生成过程的平均值之间的差异（偏差）(bias)

3. Classification problems can be solved using a regression-type approach with the logistic regression algorithm. Explain the principles of binary logistic regression, how it can be extended to multi-class problems, and what advantages and disadvantages it has over other methods.
   分类问题可以使用逻辑回归算法的回归型方法来解决。 解释二元逻辑回归的原理，如何将其扩展到多类问题，以及它相对于其他方法有哪些优点和缺点。

   The key points are:

    - Assume one can construct a function that compute the probability of a data point being in one of two classes.
      假设可以构造一个函数来计算数据点属于两个类别之一的概率。
    - Construct the log-odds of a point being in one of the classes.
      构造某一类中某个点的对数赔率。
    - Fit the log-odds of the binary decision with a linear model.
      使用线性模型拟合二元决策的对数赔率。
    - Form the joint probability density function of the data.
      形成数据的联合概率密度函数。
    - Rewrite the joint PDF in terms of the fitted model.
      根据拟合模型重写联合 PDF。
    - Maximise the likelihood to find optimal model parameters.
      最大化找到最佳模型参数的可能性。
    - Construct explicit decision rule.
      构建明确的决策规则。
    - Handle multi-class case by pivoting against one class.
      通过针对一类来处理多类案例。
    - Does not require the form of the PDF to be known.
      不需要知道 PDF 的形式。
    - Very good out-of-the box.
      开箱即用非常好。
    - Can be easily controlled by regularisation.
      可以通过正则化轻松控制。
    - Tends to be better with more data.
      更多数据往往会更好。

Question 3

1. sketch an example of a 2-dimensional dataset containing three classes of data point (unlabelled) that could not be separated by k-means clustering using Euclidean dis- tance, indicating on your sketch how k-means would incorrectly partition the data
   绘制一个二维数据集的示例，其中包含三类数据点（未标记），这些数据点无法通过使用欧几里德距离的 k 均值聚类进行分离，并在草图上指示 k 均值如何错误地划分数据

    - Any reasonable example that cannot be separated by three straight lines will suffice [3 marks]
    - The incorrect partitions should be approximately midway between the cluster centroids [2 marks].

2. researcher in the School of Chemistry has asked for your help. They have been collecting samples of water from different places around the world and have been trying to measure what is in the samples to understand the effects of environmental pollution. They have analysed all of the samples using a technique called mass spectrometry, which measures the number of molecules of a particular mass, for a range of different masses. For each sample, this produces a histogram that shows how many molecules of each mass were in the sample. This histogram can be represented as a vector, where each component corresponds to a mass value, and the value of the component is the number of molecules of that mass. The instrument used to obtain this data can measure the number of molecules at each of 1.2 million different mass values.
   The researcher has samples from 10,000 different locations and wishes to separate them into groups of similar water composition. Suggest how you would do this, highlighting any potential problems you might encounter, and how you would solve them.
   化学学院的研究员请求你的帮助。 他们一直在从世界各地收集水样本，并试图测量样本中的成分，以了解环境污染的影响。 他们使用一种称为质谱法的技术分析了所有样品，该技术可以测量一系列不同质量的特定质量的分子数量。 对于每个样品，这都会生成一个直方图，显示样品中每种质量的分子数量。 该直方图可以表示为一个向量，其中每个分量对应一个质量值，该分量的值是该质量的分子数。 用于获取该数据的仪器可以测量 120 万个不同质量值中每一个的分子数量。
   研究人员从 10,000 个不同地点收集了样本，并希望将它们分成具有相似水成分的组。 建议您将如何做到这一点，突出显示您可能遇到的任何潜在问题以及您将如何解决这些问题。

    - This is a clustering problem so could be solved with k-means or hierarchical clustering. [3 marks]
    - Issues will include high dimensionality causing slow clustering (solved by dimen-sionality reduction); lack of knowledge of number of clusters, which will need to be solved by cross-validation; the possibility that the sample might not be representative of the data. [4 marks]

3. The researcher uses another technique to identify 12 distinct types from 500 ran-domly chosen samples. Suggest how you might use this information to improve your results from part (b), again highlighting any potential problems you might encounter, and how you would solve them. 研究人员使用另一种技术从 500 个样本中识别出 12 种不同的类型。严格选择的样本。 建议您如何使用此信息来改进您的(b) 部分的结果，再次强调您可能遇到的任何潜在问题，以及你将如何解决它们。

    - his is a very challenging open-ended question that is designed to test student’s ability to think outside of the taught material. The key idea here is that the problem is now semi-supervised. Possible approaches could include:
        - Train a classifier on the labelled points and use this to classify the unlabelled points.
        - Use the labelled points to seed k-means clustering.

   One potential problem is that there may be more than twelve clusters in the data. Credit will be awarded for imaginative and self-consistent solutions that demonstrate insight into and understanding of the question. [8 marks]

   - 这是一个非常具有挑战性的开放式问题，旨在测试学生在教学材料之外思考的能力。 这里的关键思想是问题现在是半监督的。 可能的方法包括：
     - 在标记点上训练分类器，并使用它对未标记点进行分类。
     - 使用标记点来种子 k 均值聚类。

   一个潜在的问题是数据中可能有超过十二个簇。 富有想象力且自洽的解决方案将获得学分，这些解决方案表现出对问题的洞察力和理解。 [8 分]

20_resit

Question 1

1. In the notation used in the lectures, the queantities needed to solve a univariate unregularised least squares regression problem are:
   - The vector of independent variables x with components $\{x_i\}_{i=1}^N$
   - The vector of independent variables y with components $\{y_i\}_{i=1}^N$
   - The vector of independent variables w with components $\{w_i\}_{i=1}^M$
   - The basis states $\{\phi_i(x)\}_{i=1}^M$ Explain how to construct the normal equations for unregularised regression from these quantities. You do not need to derive the normal equations from first principles 解释如何根据这些量构建非正则回归的正规方程。 您不需要从第一原理推导出正规方程
   - The normal equations are $\phi^T\phi w - \phi^T y = 0$, where the components of ϕ are $\phi_{ij} = \phi_j(x_i)$
2. Explain the meaning of bias and variance in the context of a regression problem, illustrating your answer with approprivate diagrams.
   - Bias: ability of model to represent the data(low bias is good)
   - Variance: sensitivity of model to noise in the training data
   - Low bias typically requires complex model which is prone to being sensitive to the data(high variance)
   - A suitable diagram that illustrates this should be appropriately credited
3. Explain the principle of regularisation and write down the general expression for the regularised least-squares loss function. Give two examples of regularisation functions and explain their effect
   - Add term to loss function to penalise solutions with a certain structure or chracteristic and therefore encourage solutions that do not have this characteristic

   $$L(w) = ||y - f(x, w)||^2 + \lambda R(w)$$

   - L_2 regularisation penalises large values of the model parameters and therefore encourages "shrinkage"
   - L_1 regularisation also penalises large values of the model parameters and encourages "shrinkage", but also tends to encourage sparsity in the parameter vector

Question 2

1. A decision stump is a decision tree containing only one split on the most informative variable. Using the principle of maximising information gain, determine which variable should be used to form a decision stump for the data shown in the table below:

x0	x1	x2	y
0	1	1	0
0	0	1	1
1	1	0	0
1	0	1	1

   - A sharp-eyed student will notice that the dataset can be split on x1 to give homogeneous subgroups. A more pedestrian but just as acceptable approach is to compute the information gain for each variable. Both are acceptable
2. Explain the random forest algorithm for classification
   - The key points are:
     - Multiple decision trees
       - 多个决策树
     - Each tree trained on random subspace
       - 每棵树都在随机子空间上训练
     - Bagging: each tree trained on random sample from data(with replacement)
       - Bagging：每棵树都根据数据中的随机样本进行训练（有替换）
     - Decisions are taken by the majority vote of the trees
       - 决策由树木的多数票做出
3. A labelled dataset contains 500 samples, each of which is from a 5-dimensional space. It is known that there are three (3) classes of data(A,B,C) in this data set and each sample is drawn from one of those classes. The number of training points in each of the classes is A:50; B: 250; C:200. The classes are known to not be fully separable by three hyperplanes.
   Explain how you would choose an algorithm to classify this dataset, what difficulties may be encountered, and how you would overcome them.
   - The main points here are:
     - Low-d space so dimensionality reduction not necessary
       - 低维空间，因此不需要降维
     - Not separable by hyperplanes therefore no point considering LDA
       - 不可被超平面分离，因此没有必要考虑 LDA
     - Classes are unbalanced so need to take care with a majority voting technique like k-nearest neighbours
       - 类不平衡，因此需要注意 k 最近邻等多数投票技术
     - It will be necessary to cross0validate a range of techniques on this data as a priori selection looks difficult
       - 有必要对该数据进行一系列技术的交叉验证，因为先验选择看起来很困难

Question 3

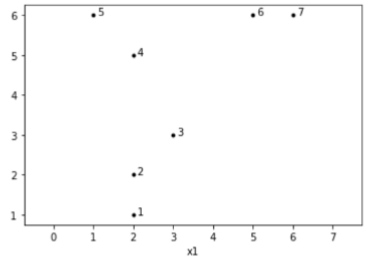
1. The Johnson-Lindenstrauss lemma can be stated as:

   $$1 - \epsilon \leq \frac{||f(x_1) - f(x_2)||^2}{||x_1 - x_2||^2} \leq 1 + \epsilon$$

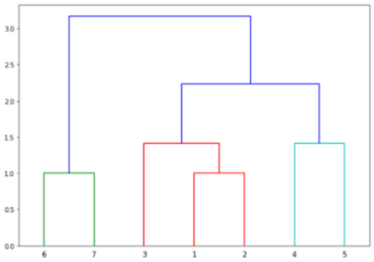
   where $x_1, x_2 \in \mathbb{R}^M; f: \mathbb{R}^M \Rightarrow \mathbb{R}; 0 < \epsilon < 1 \; and \; K < M$
   Explain the implications of this lemma and their relevance to machine learning
   - J-L is a statement that there exists a mapping f: $\mathbb{R}^M \Rightarrow \mathbb{R}^K$ from a high-dimensional
2. The table below contains a set of data with two variables. Each column contains on datapoint. Sketch the dendrogram for agglomerative hierarchical clustering using single-linkage on this dataset

x0	2.0	2.0	3.0	2.0	1.0	5.0	6.0
x1	1.0	2.0	3.0	5.0	6.0	6.0	6.0

- The data is plotted below:



from which the dendropgram can be easily derived as



3. A common modification to the k-nearest neighbours algorithms is the weighted k-nearest neighbours algorithm
   - Describe how the weighted k-nearest neighbours algorithm works
     - The pseudocode is:

```
for each training point p
    for each class c
        compute the mean distance between p and the points in C assign p to the class with the minimum mean distance
```

   - Sketch one example of a situation in which this menthod will give rise to an incorrect decision. Explain your reasoning
     - This can happen when the two classes are close and one of them is very compact. Any feasible example will suffice here as long as the reasoning is sound.

21_Main

Question 1

1. Explain what is meant by “dimensionality reduction” and why it is sometimes necessary
   - Finding a basis(coordinate system) in which the data can be represented in terms of a reduced number of coordinates without loss of significant information. It is neccessary because of the curse of dimensionality which leads to problematic phenomena such as convergence of distances and ultra-sparse sampling, and somethimes also because it allows the data size to be reduced
     - 找到一个基础（坐标系），其中数据可以用减少的坐标数来表示，而不会丢失重要信息。 这是必要的，因为维数灾难会导致距离收敛和超稀疏采样等问题现象，有时还因为它可以减少数据大小
2. Consider the following dataset of four sample points $\{x^{(i)}\}_{i=1}^{4}$ with $x^{(i)} \in \mathbb{R}^2 \forall i$:

$$X = \begin{pmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{pmatrix}$$

   Explain how to calculate the principal components of this dataset, outlining each step and performing all calculations up to (but not including) the computation of eigenvectors and eigenvalues.
   - From data matrix X subtract column means to form

$$X' = \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{pmatrix}$$

   - Form the covariance matrix

$$C = X^T X = \begin{pmatrix} 10 & 6 \\ 6 & 10 \end{pmatrix}$$

   - Find eigenvalues/vectors - NOT needed as we did not do this in class
   - Ordered eigenvalues correspond to the variance of the data projected onto the corresponding eigenvector. Usually aim to remove directions with small variance as not informative
3. What does principal component analysis(PCA) tell you about the nature of a multivariate dataset? Explain how it can be used for dimensionality reduction?
   - Aligns coordinates with natural directions in the data in order of decreasing variance. Directions in which the data does not vary can be considered to be unimportant and thus removed, reducing the dimensionality
     - 按照方差递减的顺序将坐标与数据中的自然方向对齐。 数据不变的方向可以被认为是不重要的，从而被删除，降低维度
4. What are the limitations of PCA and what other dimensionality reduction techniques may be used instead?
   - PCA can be costly to perform, and is a strictly linear technique. RP is very cheap but it is also linear and inter pretation is difficult
     - PCA 的执行成本可能很高，并且是一种严格线性的技术。 RP 非常便宜，但它也是线性的，解释起来很困难
5. You are given a dataset consisting of 100 measurements, each of which has 10 variables. The eigenvalues of the covariance matrix are shown in the following table:

Eigenvalue number	1	2	3	4	5	6	7	8	9	10
Eigenvalue	1382.0	508.4	187.0	68.8	25.3	9.3	3.4	1.3	0.46	0.17

   What can you say about the underlying nature of this dataset?
   - 95% variance is in first three dimensions, with 99% in the first five dimensions. Likely that true dimensionality of data is in this region. Unlikely that data is genuinely 10-dimensional
     - 95% 的方差发生在前三个维度，99% 的方差发生在前五个维度。 数据的真实维度很可能就在这个区域。 数据不太可能是真正的 10 维

Question 2

1. Consider the Soft Margin Support Vector Machine learnt in Lecture 4e. Consider also that C = 100 and that we are adopting a linear kernel, i.e. $k(x^{(i)}, x^{(j)}) = {x^{(i)}}^T x^{(i)}$. Assume an illustrative binary classification problem with the following training examples:

$$x^{(1)} = (0.3, 0.3)^T, y^{(1)} = 1x^{(2)} = (0.6, 0.6)^T, y^{(2)} = 1x^{(3)} = (0.6, 0.3)^T, y^{(3)} = -1x^{(4)} = (0.9, 0.6)^T, y^{(4)} = -1$$

   Which of the Lagrange multipliers below is(are) a plausible solution(s) for this problem?
   1. $a^{(1)} = 0, a^{(2)} = 2, a^{(3)} = 2, a^{(4)} = 10$

   2. $a^{(1)} = 0, a^{(2)} = 44, a^{(3)} = 22, a^{(4)} = 22$
   3. $a^{(1)} = 0, a^{(2)} = 200, a^{(3)} = 100, a^{(4)} = 100$

   - Items (i) and (iii) do not satisfy the constraints of the dual representation of the problem. In particular, (i) does not satisfy $\sum_{n=1}^{N} a^{(i)}y^{(i)} = 0$, whereas (iii) does not satisfy $0 \leq a^{(i)} \leq C$. Only (ii) satisfies all the constraints, and is thus plausible

2. Consider a binary classification problem where around 5% of the training examples are likely to have their labels incorrectly assigned(i.r.,assigned as -1 when the true label was +1, and vice-versa). Which value of k for k-Nearest Neighbours is likely to be better suited for this problem: k = 1 or k = 3?

   - The value k = 3 is likely to be better suited for this problem. This is because adopting k = 1 will cause the classifier to be sensitive to noise, whereas k = 3 is likely to reduce a bit of this sensitivity

3. Consider a binary classification problem where you wish to predict whether a piece of machinery is likely to contain a defect. For this problem, 0.5% of the training examples belong to the defective class, whereas 99.5% belong to the non-defective class. When adopting Naive Bayes for this prolem, the non-defective class may almost always be the predicted class, even when the true class is the defective class. Explain why and propose a method to alleviate this issue.

   - his is going to happen because the class-conditional probabilities are multiplied by the prior probability of the class when computing the probability of a given example belonging to this class. This prior probability is too low for the minority class, bringing the whole probability of the example belonging to this class down and making it rather unlikely that the classifier will predict the minority class
   - Possible ways to alleviate this issue include forcing the prior probability of the classes to be 50%, no matter the actual prior probability of the class; or oversampling examples of the minority class; or undersampling examples of the majority class

Question 3

1. In a small universe of five web pages, one page has a PageRank of 0.4. What does this tell us about this page?

   - the page is more important than average, as measured by the number of links that point to it. If all pages were equally important, the pagerank would be 0.2. This page is therefore much more heavily linked to than the other pages.

2. Compare and contrast the TF-IDF and word2vec approaches to document vectorisation. You should explain the essential principles of each method, and highlight their respective advantages and disadvantages

   - TF-IDF uses pure word frequency information locally (TF) and globall (IDF). Does not capture any semantic information, in particular about word order. Can be applied to relatively small data. Documents and terms can be added on-the-fly. Interpretable representation.
   - word2vec learns the semantic context of words as it is trained to predict either a missing word (bas of words) or a word's context (skip-gram). Generally needs large corpus. Cannot be retrained on-the-fly. Can learn non-linear relationships within the training set. Representation not interpretable.

3. One possible approach to searching a large linked set of documents is to combine a measure of document similarity such as TF-IDF similarity with a measure of a page's importance such as that provided by PageRank. Suggest three ways in which this could be done and discuss the advantages and disadvantages of each of them.

   - the idea that these two concepts can be combined for high quality information retrival was discussed in lectures, but now how it could actually be done. A number of possible creative solutions are possible here.
   - Potential approaches that might be discussed are:
     - Compute the document similarity vs all document, multiply by the Page rank and sort. This is simple to implements but computationally expensive for very large document sets. It also implicitly means that it's very hard to compensate for a low Page rank which could lead to less relevant documents being returned.
     - As above, but adding the two terms. This is more appealing because it allows for a strong similarity to overcome a low Page rank to some extent.
     - Sort the documents by Page rank and then compute the similarity of documents against some top fraction of the sorted documents. This is very efficient be- cause the Page rank is precomputed, but it will give very poor results because there is no guarantee that relevant documents will have a high Page rank.
     - Compute the similarity of the query against all documents, select the top matches, and then order these by Page Rank. This is likely to give the best matches, but is computationally very expensive

21_Resit

Question 1

1. Explain the purpose of the k-means algorithm and how it works.

   - Find groups of points in a dataset.
   - Choose the number of clusters to find.
   - Randomly allocate points to the clusters.
   - Compute thte centroid(mean) of the clusts.
   - Re-allocate points to the cluster with the closest centroid.
   - Repeat until clusters are stable

2. Give two examples of distance(also known as similarity) metrics commonly used in clustering algorithms and explain how they affect the result obtained.
   给出聚类算法中常用的距离（也称为相似性）度量的两个示例，并解释它们如何影响所获得的结果。

   - Euclidean distance groups points that are close in terms of straight-line distance.
   - Cosine distance groups oints together that are in the same direction from the origin

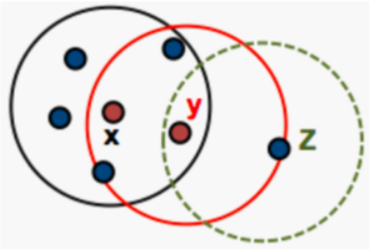
3. Explain when you would use k-means clustering and when you would use hierarchical clustering

   - In hierarchical clustering, no assumpition about the number of clusters is made whereas in k-means clustering, the number of clusters to be made are specified beforehad. What is useful is that if unaware about the number of clusters to be formed, use hierarchical clustering to determine the number and then use k-means clustering to make more stable clusters as hierarchical clustering is a single-pass exercise whereas k-means is an iterative process

4. A dataset $X = \{0, 2, 4, 6, 24, 26\}$ consists of six one-dimensional data points. The k-means clustering algorithm is initialized with 2 cluster centres at $c_1 = 3$ and $c_2 = 4$. What are the values of c_1 and c_2 after one iteration of k-means?

   - Iteration 1:
     - for c_1
       - 3, 1, 1, 3, 21, 23
     - for c_2
       - 4, 2, 0, 2, 20, 22
     - So $c_1 = \{0, 2\}$, $c_2 = \{4, 6, 24, 26\}$
     - So new $c_1 = 1$, new $c_2 = 15$
   - Iteration 2:
     - for c_1
       - 1, 1, 3, 5, 23, 25
     - for c_2
       - 15, 13, 11, 9, 9, 11
     - So $c_1 = \{0, 2, 4, 6\}$, $c_2 = \{24, 26\}$
     - So new $c_1 = 3$, new $c_2 = 25$
   - Iteration 3:
     - for c_1
       - 1, 1, 3, 5, 23, 25
     - for c_2
       - 15, 13, 11, 9, 9, 11
       - So $c_1 = \{0, 2, 4, 6\}$, $c_2 = \{24, 26\}$
       - No changes, iteration ends

5. In density based clustering, each data point is categorised as being a 'core' point, a 'border' point or a 'noise' point. The figure below shows multiple data points, three of which are labelled as x, y, and z. The circles represent the Eps-Neighbourhoods of the three labeeled points and the parameter MinPts = 6. Identify whehter each of the points (x,y,z) is a 'core' point, a 'border' point ore a 'noise' point



   - A point is a core point if it has more than a specified number of points (MinPts) within Eps These are points that are at the interior of a cluster.
   - A border point has fewer than MinPts within Eps, but is in the neighbourhood of a core point.
   - A noise point is any point that is not a core point nor a border point.
   - Therefore: x = core, y = border, z = noise

Question 2

1. Consider the follwing optimisation problem corresponding to Soft Margin Support Vector Machines:

$$\arg\min_{w,b,\xi}\{\frac{1}{2}||w||^2 + C\sum_{n=1}^{N}\xi^{(n)}\}$$

   subject to

$$y^{(n)}f(x^{(n)}) \geq 1 - \xi^{(n)}, \forall n \in \{1, 2, ..., N\}$$

   where w are the hyperplane parameters, b is the bias, ξ are the slack variables, $(x^{(n)}, y^{(n)})$ is the training example n, and N is the number of training examples. Should the constant C be positive or negative?

   - It should be positive.
   - As this is a minimisation problem and we want to reduce the amount of slack $\xi^{(n)}$
   - So that we don't have too many examples too far away from the correct side of the margin
   - Had it been negative, we would be encouraging to increase the amount of slack instead of reducing it

2. Consider the k-Nearest Neighbour algorithm learnt in Lecture 3b, applied to classification problems. In this algorithm, all k nearest neighbours contribute equally to the prediction of a given example. One may wish that examples closer to the example being predicted contribute more towards such prediction. Propose an alteration to the k-Nearest Neighbour algorithm that satistfies this requirement. Explain how this alteration works.

   - One can use the weighted majority vote instead of a simple majority vote.
   - The weight can be set to the inverse of the Educlidean distance

Question 3

1. You are given the following three documents.

   - d_1: The cat sat on the dog's mat
   - d_2: The dog chased the cat
   - d_3: The dog ate its dinner

   Stop words (the, on, its) are removed and the documents are stemmed.
   Construct the document index for these documents following stop-word removal and stemming. Explain why this data structure is useful.

   - The document index is a look-up table that is indexed by the document vocabulary. For each term in the vocabulary, the index contains the id of the document, and the number of occurences of that term in that document. This is useful beacuse it allws the inverse document frequency and the term frequencies to be computed easily by counting the elements in the table.

cat	sat	dog	mat	chase	ate	dinner
(1, 1)	(1,1)	(1,1)	(1,1)	(2,1)	(3,1)	(3,1)
(2, 1)		(2,1)				
		(3,1)				

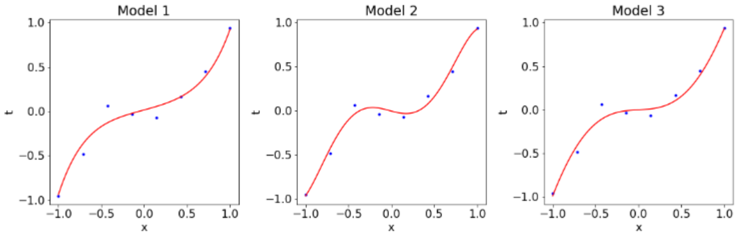
2. Compare and contrast the LSA and word2vec methods for semantic embedding.

   - LSA: based on term frequency only. Topics(embedding dimensions) derived from correlations across a corpus. Models documents as linear combinations of topics. Can work well with small corpora but does not scale well to large datasets. Does not include any sub-document information
   - word2vec: predictive model based on either BoW(predict missing word) or skip-gram(predict word context) approach. Takes both global and local context into account. Generally requires large corpus.

22_Main

Question 1

1. The images below show the results of fitting a dataset of N = 8 points(blue points) with a polynomial $f(x, w) = \sum_{i=0}^{5} w_i x^i$. The line of best fit in each case is shown as a solid red line. Each fit was generated by minimising a least squares loss function with different regularisation applied.



   The model parameters for each of the fits are.

	w0	w1	w2	w3	w4	w5
Model 1	0.015	0.265	0.018	0.378	-0.042	0.287
Model 2	-0.003	-0.261	0.170	2.389	-0.186	-1.184
Model 3	0.000	0.052	0.000	1.094	-0.027	-0.187

   State what regularisation you think was used for each fit and explain your reasoning.

   - Model 1: L2-regularised linear regression. The clue to this is in the shrinkage of the parameter vector compare to model 2.

        - 模型 1：L2 正则化线性回归。 线索在于与模型 2 相比参数向量的收缩。
    - Model 2: unregularised linear regression. The clues to this are the presence of high order terms of similar magnitude to the low order terms.
        - 模型 2：非正则线性回归。 对此的线索是存在与低阶项大小相似的高阶项。
    - Model 3: L1-regularised linear regression. The clue to this is in the sparsification of the parameter vector
        - 模型 3：L1 正则化线性回归。 线索在于参数向量的稀疏化

2. You are working on a problem involving online Bayesian Regression on data that arrives in a stream. Beginning with a Gaussian prior with mean zero and variance one, you update the posterior distribution by multiplying by the likelihood of each event in the stream as it arrives. A colleague suggests that you could improve the efficiency of this by setting a threshold below which all values of the prior are set to zero. Do you think this is a good idea? Explain your reasoning

    - It is very definitely not a good idea. Setting the prior to zero for cetain parameter values means that no matter how strong the evidence, the likelihood, which is multiplicative, can never update the posterior at those parameter values: they will stay zero even if all the evidence points to them being the correct value.

3. An instance class X contains instances $x \in X$ with twenty binary variables. A classifer L is trained to return a hypothesis from the hypothesis class containing all possible binary conjunction of three or fewer variables. Calculate an upper bound on the number of training samples needed to guarantee that the hypothesis returned by L will have a true error of no more than 10% with 80% condifence

    - Since the learner can correctly classify its training set that it is a consistent learner
    - The hardest part of this question is computation of the size of the hypothesis space. It is not $|H| = 3^{20}$(20 bits, 3 states: 0, 1, ignore) as the question asks for "all possible binary conjunctions of three or fewer variables". How of these are there?
        - All conjunctions of no variable: 1 hypothesis
        - All conjunctions of one variable: 20 variables, 2 choice for each $(A, notA)$ - 40 hypotheses
        - All conjunctions of two variable: $\binom{20}{2} = 190$, 4 choices fore each - 760 hypotheses
        - All conjunction of three variables: $\binom{20}{3} = 1140$, 8 choices for each - 9120 hypotheses
        - There are therefore $|H| = 1 + 40 + 760 + 9120 = 9921$
    - The upper bound is then given by $m \geq \frac{1}{\epsilon}(ln|H| + ln\frac{1}{\delta})$ with $\epsilon = 1$ and $\delta = 0.2$ which gives m = 108

Question 2

1. Consider the following two-dimensional data which ahs been divided into two clusters as shown:

    - Cluster 1: (2, -3), (2, 1), (3, 2)
    - Cluster 2: (-3, 1). (-3, -2)

    Given the above information, how can we determine which cluster a new point (0,0) belongs to by using k-means clustering algorithm? Show all the working/reasoning to support you answer. Describe any assumptions you may consider.

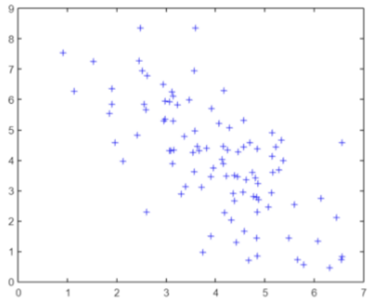
    - For this question, we need to identify that we can take two alternating positions with both of them having merits. For full credits, the answer should identify both positions and highlight the issues.
    - Position 1: we can find the distance of new point to both the cluster centres and assign it to the nearest cluster. However, this would necessitate recalculation of means after object assignment which could in turn change the cluster. First iteration is shown below:
        - As a first step, the means of cluster 1 and cluster 2 should be found which are: (2.33, 0) and (-3, -0.5). Using squared Euclidean distance, the distance of new point (0,0) to both of these cluster centres is determined. The distance is: (0, 0) to cluster 1 centre is 5.4289 and (0,0) distance to cluster 2 centre is 9.25. Hence, the point will be assigned to cluster 1. We will then recalculate the new centres which will be (1.75, 0) and (-3, -0.5), but this wouldn't alter the object assignment
    - Position 2: we can take the position that since the clusters are known as per given information and thus assigning a new point to known groups becomes a classification problem rather than clustering problem. Hence, the answer could propose(and/ or show) an algorithm like k-nn for the assignment of new point to one of the two'known classes

2. Consider that we have estimated the below mean and covariance matrix of a 2-dimensional data set during the principal component analysis process.

$$\mu = \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \sum = \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}$$

    Make a rough drawing of the point cloud for this data set, assuming that the data is generated from a multivariate Gaussian probability density function. Briefly describe your observations about his data set by discussing the influence of the variance/co-variance parameters on the point cloud shape.

    - From the mean and co-variance, the rough drawing should look similar to the one below:



    - Since there is negative covariance (-2) between the two dimensions, it's showing in the point cloud that as x-axis value increases the y-axis value decreases. Further, the spread on the x-axis is narrow than the spread on the vertical axis since the x-axis variance is lower (2) compared to the y-axis variance(4)

3. You are designing a document retrieval system in which a document q should be queried against a corpus of N linked documents $\{d_1, ..., d_N\}$. Explain how you would design such a system and in particular how you would determine:

    1. Which documents should be returned in response to a query
    2. In what order the returned documents should be presented

    - This is an open-ended and creative question with no single correct solution, but we covered TF-IDF similarity and PageRank in the lectures and I would expect these to be prominent in the submissions. The basic idea is that TF-IDF can be used to filter the documents for relevance which allows an initial filtering by content, then PR can be used to order by authority. However, we did not discuss this explicitly in lectures and so students will have to be creative and I would expect the following potential solutions to come up:
    - Order by the sum of the sim and PR
    - Order by the product of the sim and PR
    - Threshold the sim and them order by PR so as only to return docs that match the query well. This is the best option here
    - Threshold by PR and the order by sim

Question 3

Sequenctial Minimal Optimisation(SMO) is a popular algorithm used with Support Vector Machines. Different variants of this algorithm have been proposed. In the version learned in this module, whenever updating the value of a given Lagrange multiplier $a^{(j)}$ in a given iteration, the new value of $a^{(j)}$ is "clipped" based on the lowest(L) and highest(H) possible values below:

- If $y^{(i)} = y^{(j)}$
  $L = \max(0, a^{(j)} + a^{(i)} - C)$
  $H = \min(C, a^{(j)} + a^{(i)})$
- If $y^{(i)} \neq y^{(j)}$
  $L = \max(0, a^{(j)} - a^{(i)})$
  $H = \min(C, C + a^{(j)} - a^{(i)})$

where $a^{(i)}$ is another Lagrange multiplier being updated and C is a hyper parameter to control the strength of the penalty incurred by the Slack variables

1. Explain briefly what could happen if one was adopting SMO and forgot to "Clip" the values of $a^{(j)}$

    - The boundaries L and H are used to ensure that not only the box constraints are satisfied, but also the constraint $\sum_{n=1}^{N} a^{(n)}y^{(n)} = 0$. If one forgets to clip these values, these constraints may not be satisfied

  - for noting the potential violation of the box constraint and
  - For noting the potential violation of other constraints

2. Explain in detail why $H = min(C, C + a^{(j)} - a^{(i)})$ is an appropriate highest possible value for $a^{(j)}$ when $y^{(j)} \neq y^{(i)}$
  - The latter above-mentioned constraint can be rewritten as follows:

$$a^{(i)}y^{(i)} + a^{(j)}y^{(j)} = \zeta, where \sum_{n \neq i,j} a^{(n)}y^{(n)}$$

  - $y^{(j)} = -1$ and $y^{(i)} = 1$, which lead to $a^{(i)} - a^{(j)} = \zeta$
  - $y^{(j)} = 1$ and $y^{(i)} = -1$, which lead to $-a^{(i)} + a^{(j)} = \zeta$
  - If we need to clip the value of $a^{(j)}$ to a highest possible value as specified in this question, this means that the adjustment in $a^{(j)}$ is increasing the value of this Lagrange multiplier
  - In the first aforementioned case, an increase in $a^{(j)}$ will require an increase of equal amount in $a^{(i)}$ so that a^{(i)} - a^{(j)} reamains equal to ζ. Similarly, in the second aforementioned case, an increase in $a^{(j)}$ will require an increase of equal amount in $a^{(i)}$ so that $-a^{(i)} + a^{(j)}$ remains equal to ζ
  - This means that highest possible value of $a^{(j)}$ in this iteration is constrained not only by the hyperparameter C but also by the maximum possible amount by which $a^{(i)}$ can increase. Specifically, the highest possible value of $a^{(j)}$ can exceed neither C nor $a^{(j)} + D$, where D is the amount by which $a^{(i)}$ can increase
  - The maximum amount by which $a^{(i)}$ can increase is $D = C - a^{(i)}$, as any increase beyond that would mean that $a^{(i)}$ becomes larger than C, violating the box constraints.
  - Therefore, the maximum value for $a^{(j)}$ in this iteration can exceed neither C nor $a^{(j)} + (C - a^{(i)})$, which leads to $H = min(C, C + a^{(j)} - a^{(i)})$

22_Resit

Question 1

1. Consider the following objects in 2-dimensions: V (2, 10), W (2, 5), X(8, 4), Y (5, 1), Z(8, 5). Using min/single link and Manhattan distance, cluster these objects using hierarchical agglomerative clustering method. Show all the working (but no need to show the dendrogram). In addition, describe the cluster formation at height/distance 3.
  - As a first step, we will need to compute the distance matrix using the Manhattan distance which will look as below:

	V	W	X	Y	Z
V		5	12	12	11
W			7	7	6
X				6	1
Y					7
Z					

  - Next step is to successively merge objects/clusters using the min/single link. The merge will happen at heights as below:
    - @1: (X, Z), V, W, Y
    - @5: (X, Z), (V, W), Y
    - @6: ((X,Z), (V,W), Y)
  - At height 3, the cluster formation will be same as at distance 1 i.e. (X,Z), V,W,Y as the next distance after 1 at which clusters merge is distance 5.

2. Consider the following three objects in 2-dimensions:

$$X = \begin{pmatrix} 1 & 6 \\ 3 & 4 \\ 5 & 2 \end{pmatrix}$$

By following part of the principal component analysis process, estimate the covariance matrix for this data.

  - As a first step, we need to estimate mean for mean subtraction: $X = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$
  - Then we need to conduct mean subtraction:$X = \begin{pmatrix} -2 & 2 \\ 0 & 0 \\ 2 & -2 \end{pmatrix}$
  - Next step is to estimate the covariance matrix using the mean subtracted data:$\sum = \frac{1}{3}\begin{pmatrix} 8 & -8 \\ -8 & 8 \end{pmatrix}$

3. The PageRank algorithm is well-known to be the basis of Google's search engine. However, PageRank is based only on the connectivity of documents and does not take their content into account at all, and therefore cannot provide results based on a specific search term. Suggest how this could be addressed.
  - What's needed is a method for assessing the relevance of a document to the query, eg the TF-IDF similarity.
  - The goal will then be to first find the documents that are relevant to the query, then to return those documents that PageRank deems authoritative
  - Searching can therefore proceed in two stages. i) use TF-IDF similarity to compute which documents match the query (using the inverse index to speed this up), possibly using some threshold below which documents are deemed irrelevant. ii) Rank the remaining documents according to their PageRank.

Question 2

1. The regularised form of the least square loss is $L(w, \lambda) = L_{err}(w) + \lambda R(w)$ where L_{err} is the least squares loss. The regularisation term $R(w) = \alpha\|w\|_2^2 + \beta\|w\|_1$ is sometimes used in practical applications of regression.
Explain what effect this term will have on the characteristics of a model fitted using this loss, and suggest when this might be useful.
  - This is the ElasticNet loss (but is not named as such to avoid students simply looking it up). It combines the L2 and L1 regularisers and therefore simultaneously **shrinks** and **sparsifies** the model. The resulting model should therefore "select" the relevant terms by the sparsification property of the L1 loss, and then shrink over the other terms. The extent to which it does each of these is determined by α/beta. Because the L2 term makes the loss convex, and the L1 term sparsifies the solution, this is particularly useful for high dimensional problems.

2. Consider a regression problem in which we aim to predict a single dependent variable t from a single independent variable x. It is known that the true data generating function is $t = h(x) + \epsilon$, where h(x) = c, a constant, and ϵ is normally distributed with mean 0 and variance $\sigma^2 = \frac{1}{2}$.
We would like to estimate the value of c by fitting a model f (x, w) = w using Bayesian regression. Our estimate for w provides an estimate for c.
The prior distribution of w is assumed to be $p(w) \propto exp(-w^2)$.
A single data point X = (x, t) = (3, 10) is known.
    1. In the absence of data, what is E[w] (the expected value of w)?
       - $p(w) = exp(-x^2)$ is a normal distribution with $E[x] = \mu = 0$
    2. Write down the likelihood of the data point X.
       - The likelihood(with $\sigma = \frac{1}{\sqrt{2}}$) is $p(t|w) = exp[-(t - f(x, w)^2)] = exp[-(10 - w)^2]$
    3. Write down the posterior distribution of w given data point X.
       - The posterior is $p(w|t) = p(w)p(t|w) = exp[-w^2 - (10 - w)^2] = exp[-(2w^2 - 20w + 100)]$
    4. Compute the posterior estimate of w by minimising the negative log of the posterior distribution. Explain your answer.
    You may use the result that a quadratic $ax^2 + bx + c$ is minimised by $x = \frac{-b}{2a}$.
       - $-\log p(w|t) = 2w^2 - 20w + 100$ which is minimised by w = 5
       - This is not the answer we would necessarily expect: the data point implies that c ≈ 10. However, we have to include the effect of the prior. Since the prior and the likelihood have the same variance, with a single data point we end up with the average of the prior and the max likelihood estimate.

3. he following five pairs of numbers were sampled from a two-dimensional normal distribution with mean $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance $\sum = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$

x_1	2.02	-0.21	1.55	-0.05	0.81
x_2	-2.08	-1.18	-0.77	-1.15	1.32

Compute the sample mean and sample covariance, and explain the implications for learning of results of your calculations

- he data were were drawn from a two-dimensional normal distribution (the population distribution) with mean $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance $\sum = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$ and therefore a training sample drawn from this distribution should ideally have the same distribution as the population
- The mean of the sample is $\mu = \begin{pmatrix} 0.82 \\ -0.77 \end{pmatrix}$ and the covariance is $\sum = \begin{pmatrix} 0.95 & -0.21 \\ -0.21 & 1.60 \end{pmatrix}$
- These are not the same as the population statistics and this illustrate the different between the true risk and the empirical risk. It is likely that this data, were it to be used as training data for some learning algorithm, would not yield an accurate hypothesis with high probability.

Question 3

1. Logistic regression is based on the cross entropy loss function shown below (to be minimised):

$$E(\vec{w}) = -\sum_{i=1}^{N} y^i \ln p_1(\vec{x}^i, \vec{w}) + (1 - y^i) \ln(1 - p_1(\vec{x}^i, \vec{w}))$$

where w is the vector of parameters of the Logistic Regression model, $x^{(i)}$ is the vector of input variables of example i, $y^{(i)}$ is the output variable of example i, N is the number of training examples, $p_1(x^{(i)}, w) = exp(w^T x)/(1 + exp(w^T x))$ and exp is the exponential function.
Answer the following questions regarding the components shown in red of this loss function

   1. What is the effect of multiplying this equation by −1 on the training process? Justify your answer in detail.
      - the value of ln $p_1(x(i), w)$ and $ln(1 - p_1(x^{(i)}, w))$ will always be either zero or negative. The value of zero happens when the probability p_1 associated to the training example is equal to the target probability. A negative value means that it is different from the target probability, meaning that this training example is incurring some loss.
      - Multiplying by -1 will mean that any incurred loss becomes positive, i.e., it leads to a worse (larger) value for the loss function.
      - As the loss function is to be minimised, multiplying it by -1 will guide the learning process towards learning weights that assign probabilities as close as possible to the target probabilities
   2. Why are the left and right terms of the summation multiplied by $y^{(i)}$ and $(1 - y^{(i)})$, respectively? Justify your answer in detail.
      - This is necessary so that the left (right) term will contribute towards the summation only when the training example i belongs to class 1 (0)
      - If that was not the case, then the value of ln p1(x(i), w) would be summed for examples of class 0, meaning that a probability p1 of zero for examples of class 0 would be considered a bad value, leading to an increase in the loss. However, such probability value should be considered as a very good value when the example belongs to class 0. A similar issue would happen with the right term being used for examples of class 1

2. The Gaussian Kernel is a very popular kernel that is frequently used with Support Vector Machines. It is defined based on a Gaussian function, which is associated to a hyperparameter σ:

$$k(x, x^{(n)}) = e^{-\frac{||x - x^{(n)}||^2}{2\sigma^2}}$$

Explain the effects that increasing and reducing the value of σ would have on the function below, which is used to predict the output value of an example described by the input vector x:

$$f(x) = \sum_{n \in S} a^{(n)} y^{(n)} k(x, x^{(n)}) + b$$

where $a^{(n)}$ is the Lagrange multiplier associated to the support vector n, $y^{(n)}$ is the output value of the support vector n, $x^{(n)}$ is the vector of input values of the support vector n, S is the set of indexes of the support vectors and

$$b = \frac{1}{N_S} \sum_{n \in S} a^{(m)} y^{(m)} k(x^{(n)}, x^{(m)})$$

where N_S is the number of support vectors
Instructions: assume that the Lagrange multipliers associated to all support vectors always have the same value, i.e., assume that the kernel and output values are the only factors influencing f (x).

   - Larger values of σ will result in a wider Gaussian function where the similarity values retrieved by the kernel would always be very similar to each other
   - This means that the prediction given to a given example will receive a lot of influence from support vectors that are not so similar to it
   - In particular, if we use an extremely large value for σ, then the similarities between different examples will always have a very similar value, meaning that all support vectors contribute almost equally to the predictions
   - In contrast, smaller values of σ will result in a narrower Gaussian function with a higher peak, where the similarity values retrieved by the kernel would only be high when the examples given as arguments to the kernel are very similar to each other
   - Therefore, only the support vectors that are very similar to the example being predicted would provide a considerable contribution to the predictions
   - In particular, an extremely small value for σ would mean that only the closest support vector would have any meagninful effect on the predictions

23_Main

Question 1

1. The least squares error function is defined as

$$L(w) = \sum_{i=1}^{N} (t_i - f(x_i, w))$$

This function is commonly used to measure how well a function $f(x, w)$ parameterised by w fits a set of N data points $D = \{(x_i, t_i)\}_{i=1}^{N}$
The likelihood of a data point t having been generated by given a model $f(x, w)$ can be written as $p(t|f(x, w))$. Explain how, and under what assumptions, the least squares error is derived from the likelihood. You do not need to reproduce all of the mathematical steps of the derivation

   - Least square error is the function(technically the functional) that, when minimised, maximises the likelihood of the data
   - under the assumption of normally distributed iid data point
   - The steps in the derivation are:
      - Assume Gaussian likelihood function
      - Assume data points are **independent and identically distributed**(iid). Then likelihood is product of identical univariate Gaussians.
      - Maximimising the likelihood is equavalent to maximising it's log
      - The log of the product of Gaussians becomes the sum of the logs of the Gaussians
      - So maximising the likelihood is equivalent to minimising the LSE

2. Given some dataset, the expected value of the LSE L can be written as:

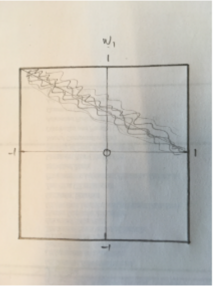
$$E[L] = \sigma^2 + var[f] + (h - E[f])^2$$

where σ^2 is the variance of the data, f is the estimated fit, and h is the ture data generating function. Explain the terms in this expression and its relevance for learning.

   - This is the bias-variance decomposition. It is the expected value of the least squares error expressed as the sum of three term:
      - The noise in the data
      - The variance of the estimator
      - The difference (bias) between the true function and the expected value of the estimator
   - It implies that for a given goodness-of-fit, there is a trade-off between the bias and variance of the estimator, with lower bias models necessarily having a higher variance for the same expected LSE and vice-versa. This places a fundamental limit on the generalisability of a learned model, because simple models will tend to have high bias, and complex models high variance.

3. Given the data point(2,1), sketch a diagram of the likelihood in parameter space that this data point was generated by functions of the form $f(x, w) = w_0 + w_1 x$. Your sketch should cover the domain $\{w_0, w_1\} \in [-1, 1]$

   - We are looking here for the set of lines that pass near to the point having a higher probability, and those that do not having alower probability. Those lines that pass through the point stisfy $w_0 + 2w_1 = 1$ or $w_1 = (1 - w_0)/2$ and so this defines the "ridge" of maximum likelihood. The sketch should therefore look something like the following:



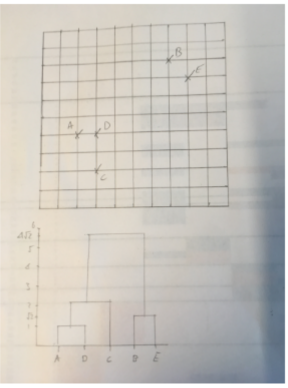
Quesiton 2

1. Cluster the data in the table below using hierarchical clustering with Euclidean distance and signle likage, and draw the dendrogram.

Label	A	B	C	D	E
Coordinates	(4,2)	(7,8)	(3,2)	(3,4)	(8,7)

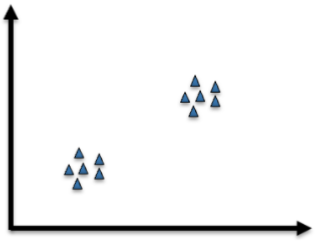
	A	B	C	D	E
A		$\sqrt{45}$	$\sqrt{1}$	$\sqrt{5}$	$\sqrt{41}$
B			$\sqrt{52}$	$\sqrt{32}$	$\sqrt{2}$
C				$\sqrt{4}$	$\sqrt{50}$
D					$\sqrt{34}$
E					

   - @$\sqrt{1}$: (A,C), B, E, D
   - @$\sqrt{2}$: (A,C), (B,E), D
   - @$\sqrt{4}$: (A,C,D),(B,E)

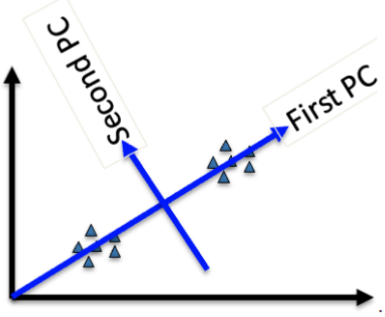


   - @$\sqrt{32}$: ((A,C,D),(B,E))

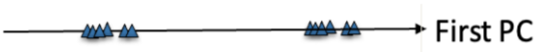
2. The graph below shows a two dimensional dataset.



   1. Reproduce the plot and draw the first and second principal components. Your drawing does not need to be completely accurate but should capture the key features. Explain your resoning.
      - The first PC should align with the direction of greatest variation in the data which in this case runs roughly through the centres of the two visible "clusters".
      - The second PC is perpendicular to this. Note that the directions of the PCs are arbitary.



   2. If you were the use PCA to reduce the dimensionality of this data to just 1 dimension, show how the points will be mapped onto the new dimension(your drawing should give the general idea of the mapping)
      - A correct sketch should show the data projected on to the first PC. The key feature is that two distinct "clusters" should remain obvious

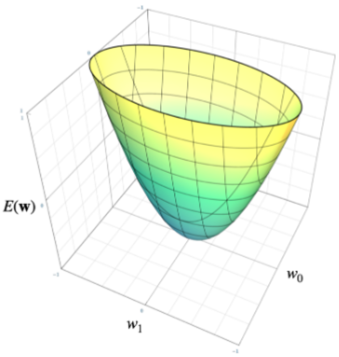


   3. Describe one way to determine how many dimensions should be kept in PCA.
      - The most common way is by computing the amount of variance explained by each principal component, and retaining those the explain about 90% of the total variance in the data, but this is highly problem dependent

3. Document vectorisation and Page Rank are both methods for ranking documents. Document vectorisation allows documents to be ranked by their similarity to a query document. Page Rank provides a way to rank a collection of linked documents by considering their authority which is derived from the connectivity of each document.
   Explain briefly how these two methods might be combined to implement a basic search engine for collection of linked documents that takes both document content and document authority into account. Briefly discuss any limitations of your approach.
   - Serveral answers are possible here.
   - The key is to realise what is required of a search engine.
   - It should return only results that are relevant, and should return those results in a sensible order.
   - Credit will be given for sensible proposals that are carefully thought-through in terms of their implications for both the quality of results and the cost
   - The optimal solution involves filtering by content similarity first, then computing authority over the subset of documents to return the most authoritative documents first

Question 3

1. Which algorithm(Gradient Descent or Iterative Reweighted Least Squares) would be better to learn the weights w_0 and w_1 of a logistic regression model for a problem with the loss function $E(w)$ below, which is an elliptic quadratic function? Justify your answer by explaining how these two algorithms would work in the context of this problem.



   - Iterative Reweighted Least Squares would be better.
   - This is because this algorithm uses a Taylor polynomial of degree two to approximate the loss function. The minimum of the Taylor polynomial is then obtained by setting the gredient to zero. As this loss function is a quadratic function, this approximation is perfect. Therefore, the minimum of the Taylor polynomial is also the minimum of this function and the algorithm would be able to find the optimum in a signle step
   - In contrastm Gradient Descent updates the weights in the direction of the steepest descent. However, as this function is elliptical, such updates are not steps that go directly to the optimum. They are likely to overshoot the optimum in the direction of the w_0 axis. Even though gradient Descent will eventually reach the optimum if a suitable learning rate is used, several steps(weight updates) would be typically necessary for that
2. Logistic regression models for binary classification can be trained by maximising the log-likelihood:

$$\ln(L(w)) = \sum_{i=1}^{N} y^{(i)} \ln p_1(x^{(i)}, w) + (1 - y^{(i)}) \ln(1 - p_1(x^{(i)}, w))$$

   where w are the weights of the logistic regression model, $y^{(i)} \in \{0, 1\}$ is the output variable of training example $i, x^{(i)} \in X$ are the input variables of training example i, X is the input space, N is the number of training examples,and $p_1(x^{(i)}, w)$ is the probability of example i to belong to class 1 given $x^{(i)}$ and w.
   How would you modify the log-likelihood function above so that it also works for problems with $M > 2$ calsses? Explain your function.
   PS: Please cereate a single log-likehood function and make sure to define any variable or symbol that is different from the ones defined above
   - $$\ln(L(w)) = \sum_{i=1}^{N} \sum_{k=1}^{M} y_k^{(i)} \ln p_k(x^{(i)}, w)$$

     where $y_k^{(i)}$ is 1 if example i belongs to class k and 0 other wise; and $p_k(x^{(i)}, w)$ is the probability of example i to belong to class k computed using w
     This function sums the log of the probability of each example to belong to its true class k. This is because the $y_k^{(i)}$ multiplying $\ln p_k(x^{(i)}, w)$ will only have value 1 for the true class to which this example belongs, resulting in the log of the probability of the example to belong to its true class being added to the summation. For all other classes, it will have value 0, resulting in zero being added to the summation.
3. Prove that the kernel below is valid kernel based on the kernel composition rules below and the fact that $x^T z$ is a valid kernel.

$$k(x, z) = 10(e^{(x^T z)})^2 + 2 + x^T z$$

   Kernel composition rules, given two valid kernels $k_1(x, z)$ and $k_2(x, z)$
   1. $k(x, z) = ck_1(x, z)$, where c > 0 is a constant
   2. $k(x, z) = f(x)k_1(x, z)f(z)$, where f (·) is any function
   3. $k(x, z) = q(k_1(x, z))$, where q(·) is a polynomial with non-negative coefficients
   4. $k(x, z) = e^{k_1(x,z)}$
   5. $k(x, z) = k_1(x, z) + k_2(x, z)$
   6. $k(x, z) = k_1(x, z)k_2(x, z)$
   - This can be proved, for example, by using the following kernel composition rules:
     - (5) to get $10(e^{(x^T z)})^2 + 2$ and $(x^T z)$
     - (3) to get $e^{(x^T z)}$ from $10(e^{(x^T z)})^2 + 2$
     - (4) to get $x^T z$ from $e^{(x^T z)}$

24_Mock

Question 1

1. Explain what is a supervised learning algorithm, including its core goal. Please give your answer formally, make use of appropriate mathematical symbols and terminology whenever relevant.
   - A supervided learning algorithm is an algorithm that takes as input a training set containing pairs of inputs and target outputs

$$T = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), ..., (x^{(N)}, y^{(N)})\}$$

   where $(x^{(i)}, y^{(i)}) \in XxY$ are drawn independently and identically distributed from a fixed albeit unkown joint probability distribution $p(x, y)$, X is the input space and Y is the output space
   It then learns a function $g : X \to Y$ with the goal of being able to predict(generalise to) unseen (test) examples of the same probability distribution $p(x, y)$
2. Answer the following questions regarding feature transformations in the context of machine learning:
   1. What is a non-linear feature transformation? Please provide a detailed definition.
      - A non-linear feature transformation is a vector function that transforms the input space of the problem using at least one non-linear function. In particular, a feature transformation is a vector function ϕ that receives an input vector x, where $x \in \mathbb{R}^d$, and $d \geq 1$ is the dimensionality of the input space of the problem. In a non-linear feature transformation, at least one of the functions$\phi(x)_i \in \phi(x)$ is a non-linear function
   2. When could it be useful to adopt a non-linear feature transformation?
      - It could be useful to adopt a non-linear transformation when the problem is non-lienar but we wish to use a linear model to solve it. In particular, a non-linear feature transformation could potentially transform the problem into a feature space where the problem is linear, such that a linear model on the feature space could be used to solve it

Question 2

1. Consider a machine learning problem with two parameters to be learned(w_1 and w_2) and the following loss function:

   $$E(w) = w_1^2 + 0.001w_2^2$$

   where $w_1 \in R$ and $w_2 \in R$

   1. Explain in detail why Gradient Descent could be inefficient to minimise this function.
      - This function has a much larger scalar being multiplied by w_1 than by w_2. If one were to plot it, it would form an elliptical bowl where the gradient in the direction of w_1 is much steeper than in the direction of w_2, $\frac{\partial g}{\partial w_1} > \frac{\partial g}{\partial w_2}$
      - As a result, given the Gradient Descent weight update rule $w = w - \eta \nabla E(w)$ and a fixed learning rate η, the size of the weight update for w_1 will be larger than that for w_2.
      - The larger weight update for w_1 may result in the algorithm jumping across the optimum in the direction of w_1, while the slower weight update for w_2 would result in the algorithm giving small steps. This could result in a long time to find the optimum.
   2. Explain in detail why standardisation of the input variables (e.g., by deducting the mean from each input variable and then dividing each input variable by the standard deviation) could be potentially helpful to improve the efficiency of Gradient Descent for this problem.
      - The reason why standardisation could help is related to the reason why the scalar multiplying w_1 is larger than the one multiplying w_2. In particular, the different sizes of the scalars could be due to a different scale for the input variables x_1 and x_2.
      - This is because the loss function is obtained by calculating the error on the training examples, which in turn is based on the predictions. For instance, assume that the predictions are given based on a function $h(x) = x_1 w_1^2 + x_2 w_2^2$. A x_1 that has a larger scale than x_2 could result in a larger scalar being multiplied by w_1 than by w_2 in the loss function.
      - Standardising the input variables will result in them being in the same scale, such that the scalars multiplying each weight would hopefully be more similar, resulting in more similar gradients. The more similar gradients would in turn reduce the problem mentioned in the previous item of this answer.

2. In the dual representation of the Support Vector Machines, it could happen that a training example is on the margin, but is associated to a Lagrange multiplier of zero. Despite being on the margin, this training example is not considered as a support vector, as it would not contribute towards the predictions made by the model. Explain in detail why an example that is on the margin could possibly have a Lagrange multiplier of zero.

   - Note the exam's annex. The dual representation is created by starting with the primal representation and then creating a penalty function g(w, b) to deal with the constraints.
   - When a training example n is on the margin, the constraint $y^{(n)}h(x^{(n)}) \geq 1$ is satisfied with the equality. As a result, the term $1 - y^{(n)}(w^T \phi(x) + b)$ in the penalty function g is equal to zero.
   - In such situation, the value of $a^{(n)}$ does not matter to maximise g(w, b). The value of $a^{(n)}$ can be any value $a^{(n)}$ ≥ 0. If $a^{(n)}$ > 0 the training example n is a support vector (it will contribute to predictions). However, as $a^{(n)}$ can take any value ≥ 0, the training example n could also be associated to a value of $a^{(n)}$ = 0, which is the case considered in this question.