FML 17-18

****

1. The main difference between these types is the level of availability of ground truth data, which is prior knowledge of what the output of the model should be for a given input.

**Supervised learning** aims to learn a function that, given a sample of data and desired outputs, approximates a function that maps inputs to outputs.

**Example:** 1. The prediction of house prices by the linear regression

2. The classification of pictures in Iris data set

**Semi-supervised learning** aims to label unlabeled data points using knowledge learned from a small number of labeled data points.

**Example:** 1. The classification of pictures but some pictures are not labeled.

**Unsupervised learning** does not have (or need) any labeled outputs, so its goal is to infer the natural structure present within a set of data points.

**Example:** 1. Use K-means clustering algorithm to learn the structure of data.

[Reference](https://towardsdatascience.com/understanding-the-different-types-of-machine-learning-models-9c47350bb68a)

**理解部分(以下部分，无需记忆，只是为了帮助理解上面的答案)**

1. Supervised learning 是有特征（feature）和标签（label）的，即便是没有标签的，机器也是可以通过特征和标签之间的关系，判断出标签。举例子理解：高考试题是在考试前就有标准答案的，在学习和做题的过程中，可以对照答案，分析问题找出方法。在高考题没有给出答案的时候，也是可以给出正确的解决。这就是监督学习。

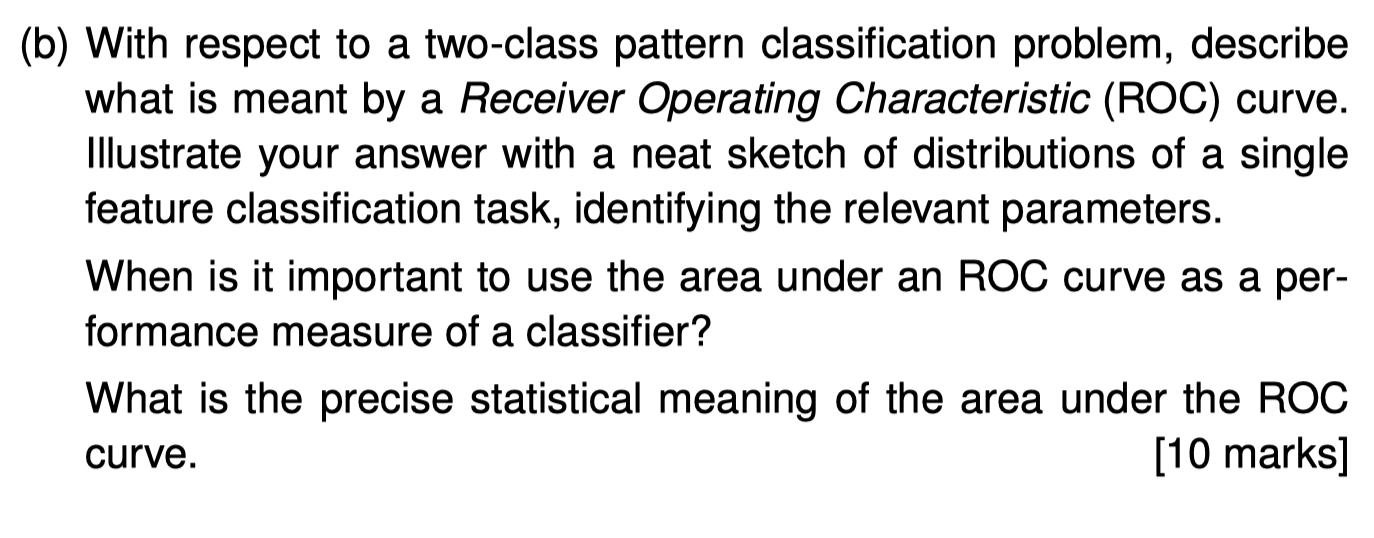
一句话概括：给定数据，预测标签。

2. Unsupervised learning 只有特征，没有标签。举例子理解：高考前的一些模拟试卷，是没有标准答案的，也就是没有参照是对还是错，但是我们还是可以根据这些问题之间的联系将语文、数学、英语分开，这个过程就叫做聚类。在只有特征，没有标签的训练数据集中，通过数据之间的内在联系和相似性将他们分成若干类。

一句话概括：给定数据，寻找隐藏的结构。

3. Semi-Supervised learning 使用的数据，一部分是标记过的，而大部分是没有标记的。和监督学习相比较，半监督学习的成本较低，但是又能达到较高的准确度。

一句话概括：部分数据有label，但大部分都没有。 [Reference](https://www.jianshu.com/p/56fe011d9bae)

****

**我们将问题拆解成几个小问题**

**1. With respect to a two-class pattern classiﬁcation problem, describe what is meant by a Receiver Operating Characteristic (ROC) curve.**

Definition: A ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. (From Wikipedia)

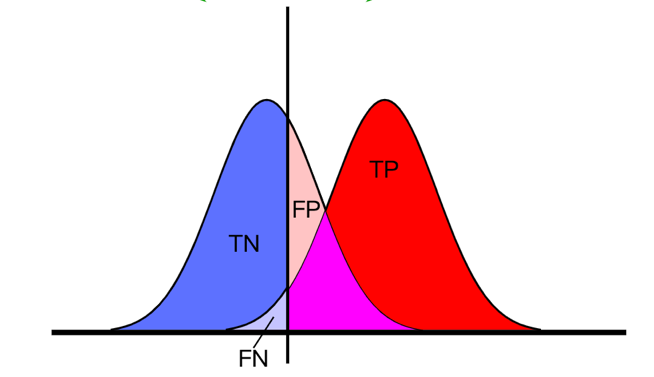
The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. (这一句也可以补充，或者辅助理解)

**2. Illustrate your answer with a neat sketch of distributions of a single feature classiﬁcation task, identifying the relevant parameters.**

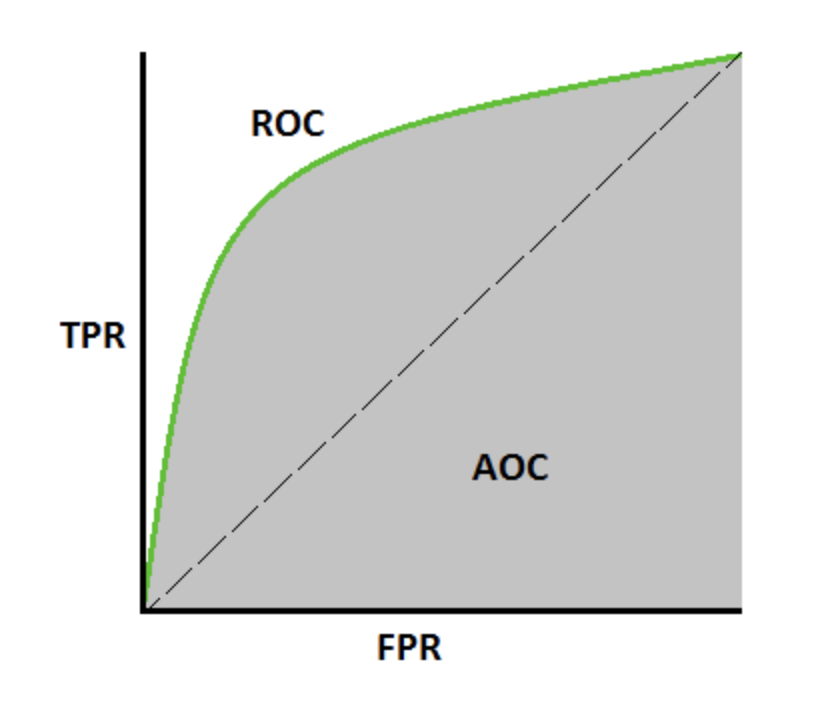
思路：

画出概率密度图像，然后解释什么叫TN,FP,TP,FN

As altering the threshold in the figure, areas of TN, FN, FP, TP, which are true positive, false negative, false positive, and true positive, also changes.



在接下来，画出ROC 曲线，开始用ROC来解释上面图中阈值的变化过程。



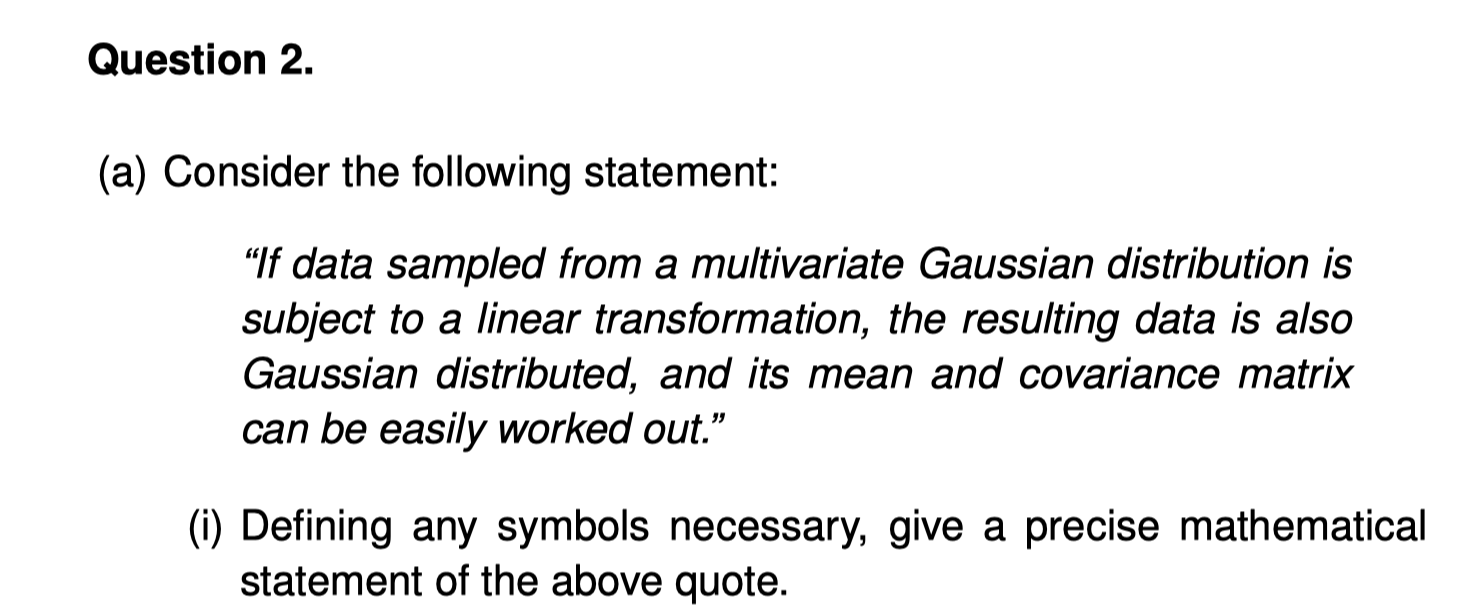
As the setting of the threshold changing, we could use FPR(false positive rate) and TPR(true positive rate) to plot ROC curve as shown in the picture.

**3. When is it important to use the area under an ROC curve as a performance measure of a classiﬁer?**

When we estimate a number of classifiers, it is important to use the area under an ROC curve as an efficient method to choose the best classifier among them.

**4. What is the precise statistical meaning of the area under the ROC curve.**

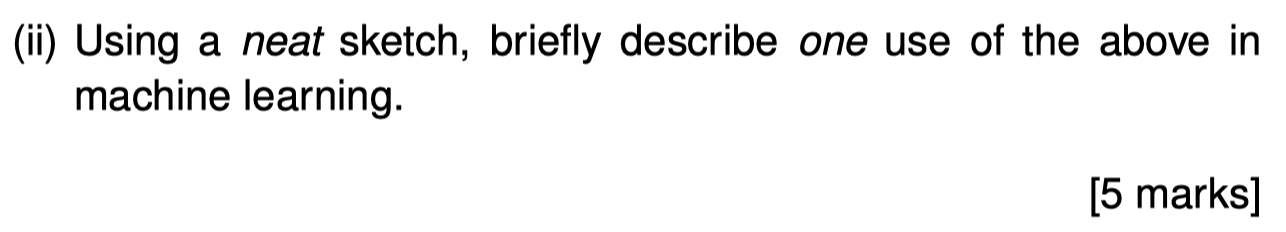
From the statistical perspective, AUC means the probability of predicting a ture positive sample rank topper than a true negative sample.

****

**（i）**

**图片包含 文字

描述已自动生成**

****

图片包含 文字

描述已自动生成

We can use this property to realize LDA. As the figure show, when we project two Gaussian distributions on the direction u, the result data are also two Gaussian distributions.

图片包含 室内

描述已自动生成

不会写，需要讨论一下

图片包含 瓶子

描述已自动生成

乍一看感觉这道题是在说“Bayes optimal classifier”,百度过后感觉这个没学过。后来看了16-17的题目，觉得这道题其实只是让我们阐述当方差和均值不同的情况下，推导贝叶斯公式。那么就有点像14-15的题目。这里只写了C1等于C2，m1不等于m2的情况。这道题的分值这么大，应该是得分情况讨论的。

注意！！这里的x是一个single continuous value。所以下面不需要带多元高斯，只需要代进去一维高斯分布的函数。这里我就不推了，细节与下面的类似，但是细节比这个简单。

图片包含 文字

描述已自动生成图片包含 文字

描述已自动生成

**Question 3**

1. **Write down the computational steps involved in the K−means clustering algorithm. Use a pseudo code format rather than a paragraph length description.**  **[5 marks]**

图片包含 屏幕截图

描述已自动生成

1. **Six one-dimensional (univariate) measurements**

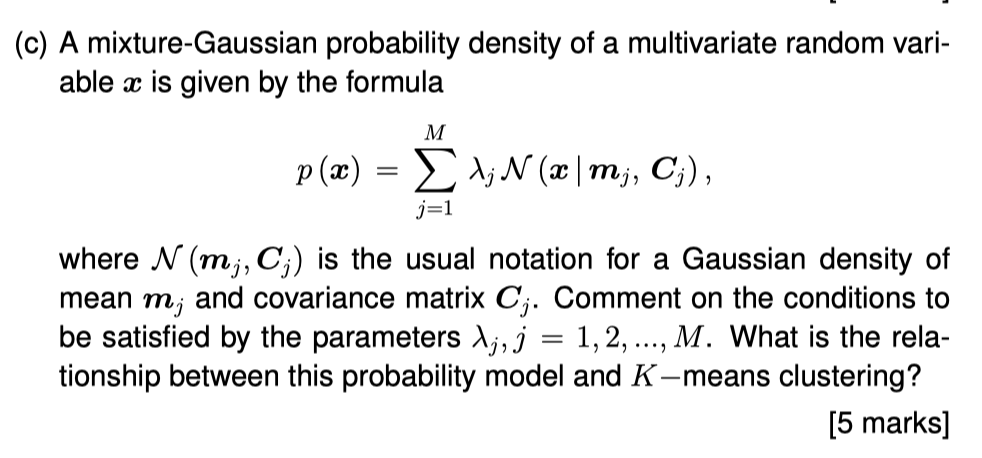
**{1, 6, 3, 8, 2, 7}**

**are to be clustered using K−means clustering into two groups (K = 2). An initial guess of the grouping has assigned the ﬁrst three items in the above sequence to one of the two clusters and the remaining three to the other cluster.**

**Starting from this setting, compute the steps involved in K−means clustering and cluster the data. [5 marks]**

**图片包含 文字

描述已自动生成**

****

Comparison of the K-means algorithm with the EM algorithm for Gaussian mixtures shows that there is a close similarity. Whereas the K-means algorithm performs a hard assignment of data points to clusters, in which each data point is associated uniquely with one cluster, the EM algorithm makes a soft assignment based on the posterior probabilities. In fact, we can derive the K-means algorithm as a particular limit of EM for Gaussian mixtures. [PRML p443, 9.3.2节]

**理解部分(不需要记忆)**

简单地说，k-means 的结果是每个数据点被分配到其中某一个 cluster 了，而 GMM 则给出这些数据点被分配到每个 cluster 的概率，又称作 soft assignment，也称为软聚类 。

GMM与KMeans区别

1）从上面的分析中我们可以看到 GMM 和 K-means 的迭代求解法其实非常相似，因此也有和 K-means 同样的问题──并不能保证总是能取到全局最优，如果运气比较差，取到不好的初始值，就有可能得到很差的结果。对于 K-means 的情况，我们通常是重复一定次数然后取最好的结果，不过 GMM 每一次迭代的计算量比 K-means 要大许多，一个更流行的做法是先用 K-means （已经重复并取最优值了）得到一个粗略的结果，然后将其作为初值（只要将 K-means 所得的 centroids 传入 gmm 函数即可），再用 GMM 进行细致迭代。

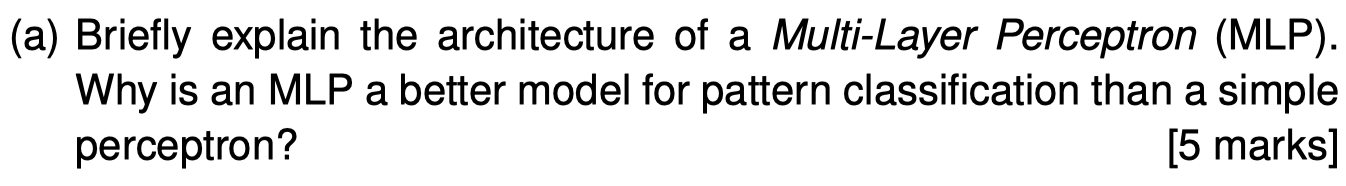
2)k-means 的结果是每个数据点被 assign 到其中某一个 cluster 了，而 GMM 则给出这些数据点被 assign 到每个 cluster 的概率，又称作 soft assignment 。

**(d)Brieﬂy describe two applications in which K−means clustering could be useful. [5 marks]**

1. **Image segmentation and compression**
2. **User Behavior Analytic**
3. **Vector Quantisation in Bag of visual words.**
4. **Recommender system**

****

**Q4基本上和14-15年的最后一题相同。唯独(b)问题不同。**

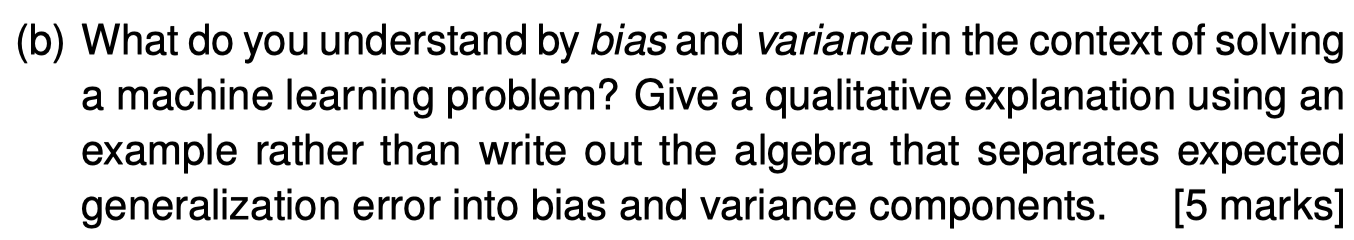
****

**先简洁的解释一下MLP的结构，红色字体是肯定要提到的**

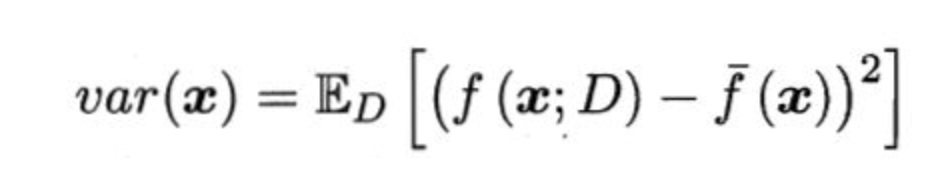
An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear [activation function](https://en.wikipedia.org/wiki/Activation_function). (From Wikipedia)

**为什么多层感知机比简单的感知机更好。那是因为多层感知机可以解决非线性问题。**

 MLP can distinguish data that is not linearly separable.

****

**虽然这道题说了不要用代数去解释bias和variance，但是我还是想把bias和variance直观的公式写下来，让我们先有个印象。**

****

**图片包含 物体

描述已自动生成**

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. [Reference](https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229)

**考场上大家可以画一个靶芯图(bulls-eye diagram)，如下图所示，浅显易懂。当然也可以看一下李宏毅机器学习的视频，**[**链接**](https://www.youtube.com/watch?v=D_S6y0Jm6dQ&list=PLJV_el3uVTsPy9oCRY30oBPNLCo89yu49&index=5)**，这一集专门讲的bias和variance的区别和联系。**

图片包含 文字

描述已自动生成

图片包含 室内

描述已自动生成

**解释一下用反向传播算法计算梯度。感觉这道题有点抽象，但是肯定需要提到chain rule。**

The backpropagation algorithm works by computing the gradient of the loss function with respect to each weight by the [chain rule](https://en.wikipedia.org/wiki/Chain_rule), computing the gradient one layer at a time, [iterating](https://en.wikipedia.org/wiki/Iteration) backwards from the last layer to avoid redundant calculations of intermediate terms in the chain rule. (From Wikipedia)

**提及两种避免过拟合的方法。学到的两种正则化方法。以及我自己加的两种。**

1. L1 regularization 图片包含 物体

   描述已自动生成
2. L2 regularization图片包含 物体

   描述已自动生成
3. Drop out
4. Cross validation

图片包含 室内

描述已自动生成

The maximum margin principle is that we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum-margin classifier. (From Wikipedia)

**然后用拉格朗日乘子法转化为目标函数（全来自于上课的PPT）**

图片包含 物体

描述已自动生成

**然后将求好的结果代入转化对偶问题**

