**FML 14-15**

图片包含 室内

描述已自动生成

1. 这道题就是解释监督学习，无监督学习，半监督学习的性质和举例

注：这道题也是17-18的原题

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)

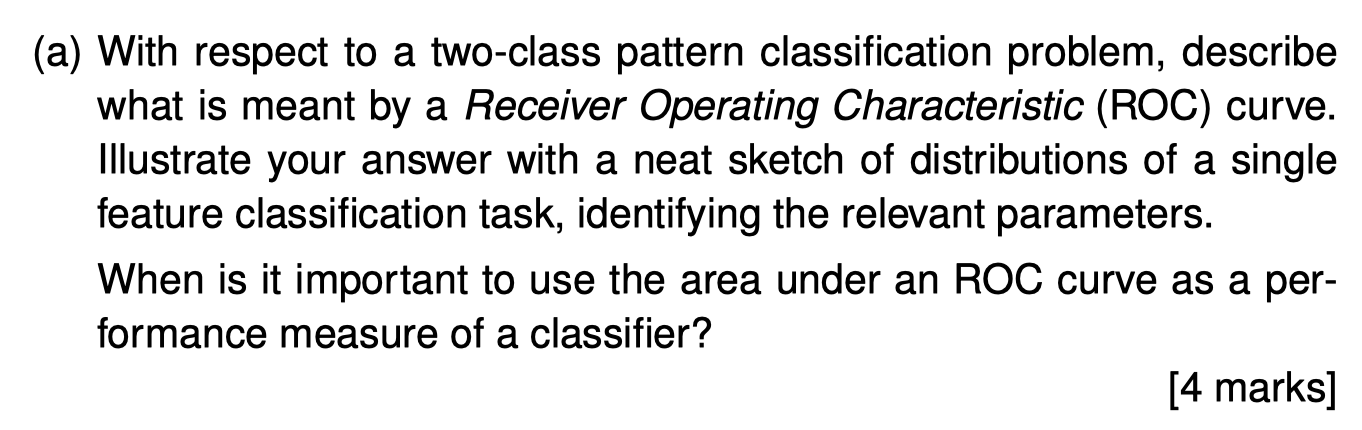
1. 这道题就是证明，在样本满足高斯分布的情况下，后验概率的结果满足sigmoid函数。

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 2**

****

注：这道题是17-18的原题。在17-18年，分值有10分，这里分值有4分。答的大差不差应该就能拿分。

思路：

**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)

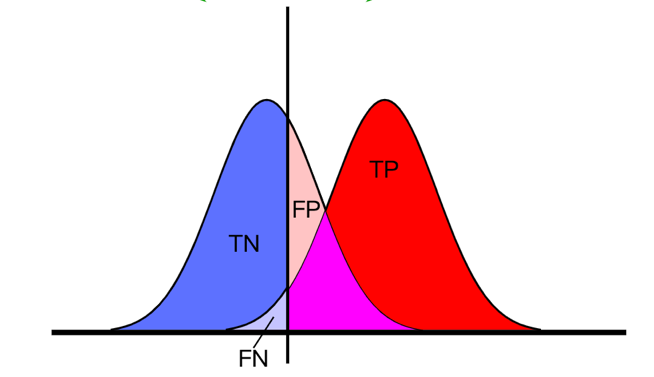
The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. (这一句也可以补充，或者辅助理解)

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

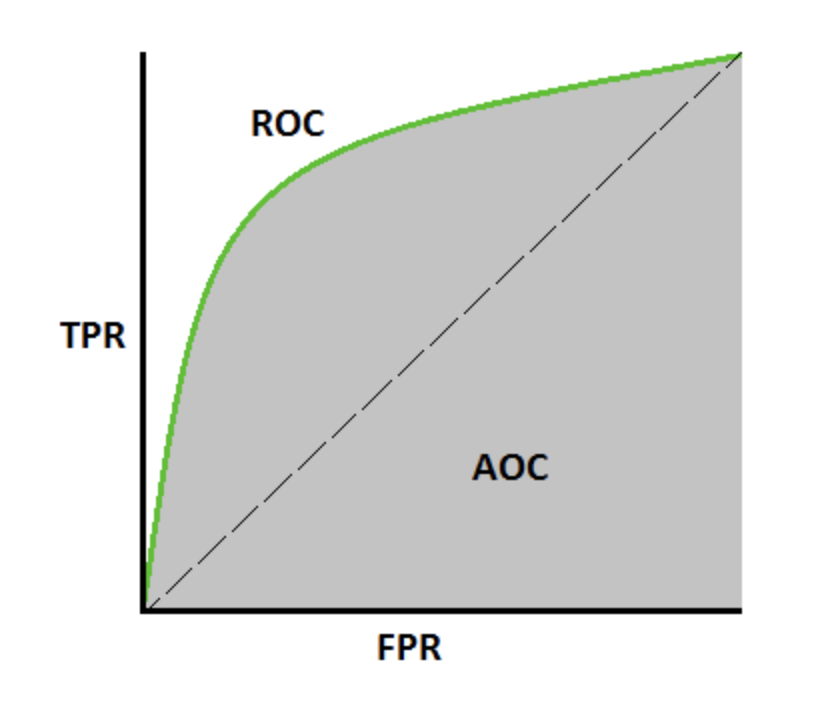
思路：

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.



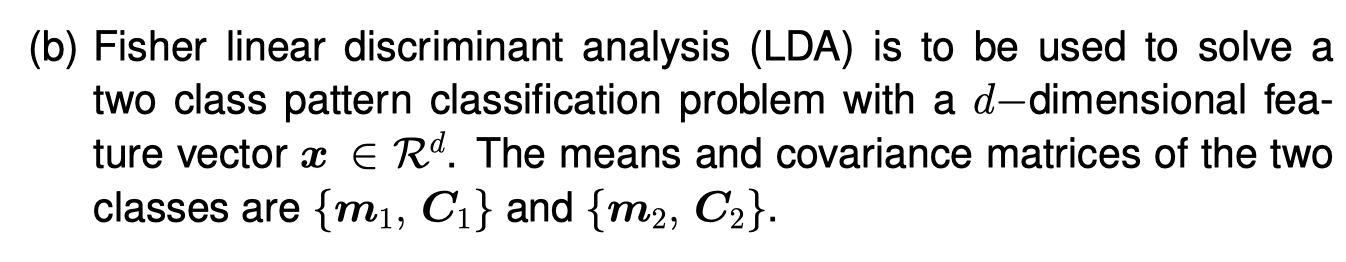
在接下来，画出ROC 曲线，开始用ROC来解释上面图中阈值的变化过程。

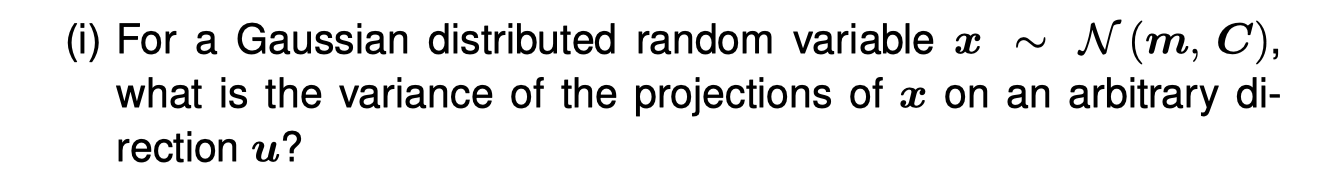


As the setting of the threshold changing, we could use FPR(false positive rate) and TPR(true positive rate) to plot ROC curve as shown in the picture.

**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.



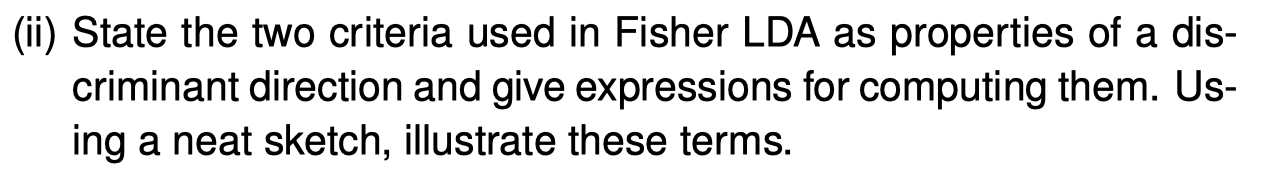


1. 这道题就是证明投影在u向量上的 方差C是多少。换句话说，我知道一批数据x满足图片包含 物体, 手表

   描述已自动生成，现在我要把它投影到u这个向量上。那么在u这个向量上，这些数据的方差是多少？这就是此小题的问题。

图片包含 文字

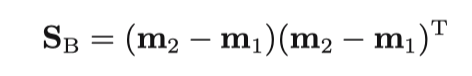
描述已自动生成



（ii） 解释下类间协方差和类内协方差

There are two criteria used in Fisher LDA, which are SB and SW. SB is the between-class covariance matrix

and SW is the total within-class covariance matrix.



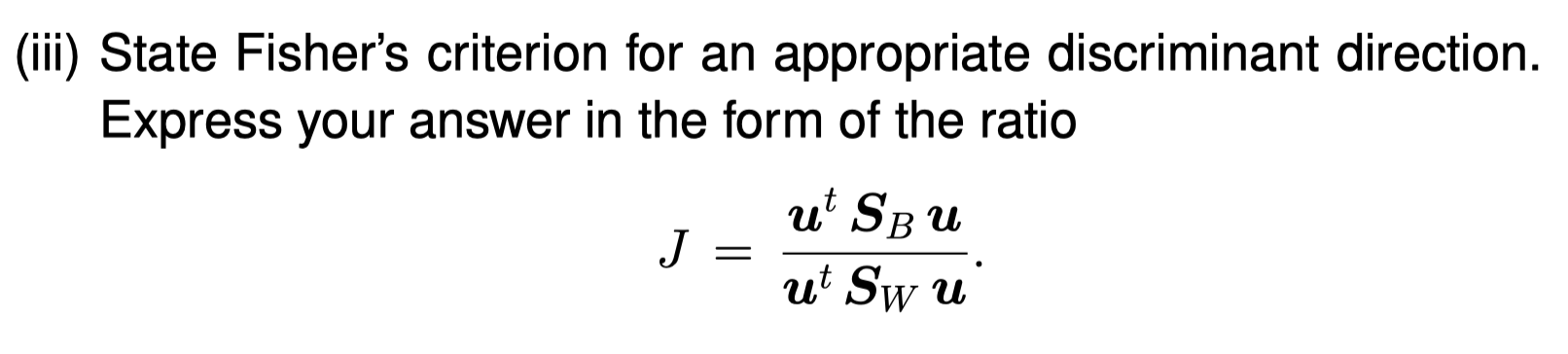
图片包含 物体

描述已自动生成

然后画一个草图，但是类间方差好表示，但是类内方差，我不知道如何表示出来。这里拿周志华书中的图片展示一下，大家考试的时候，在此图的基础上，标一标参数，意思意思。

图片包含 文字

描述已自动生成



没看懂题目，先估计一下题目是让我们阐述一下获得最佳u的条件。

图片包含 文字

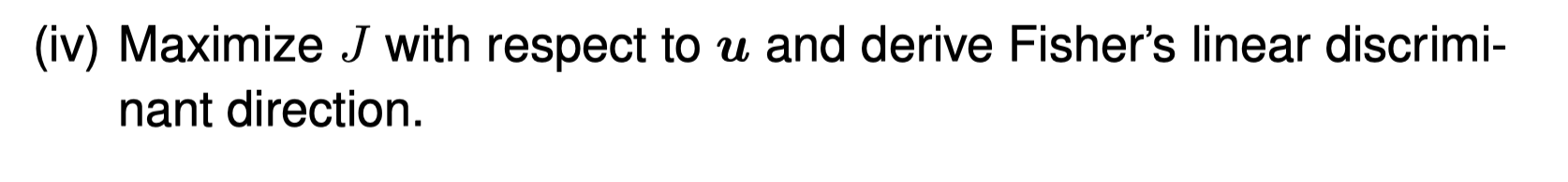
描述已自动生成

分子是两类均值的投影之间的距离(the distance between projected means of two classes),分母是被投影的类内离散程度（within class scatter of projections）。那么获得u的最大条件就是让分子尽可能的大，也就是让两类投影均值点的距离尽可能的分离（apart）,让投影的类的协方差越小。

If we want to get an appropriate discriminant direction, we should keep 图片包含 天空

描述已自动生成 as high as possible. Meanwhile, we also want 图片包含 物体, 天空, 时钟

描述已自动生成 to be as low as possible.



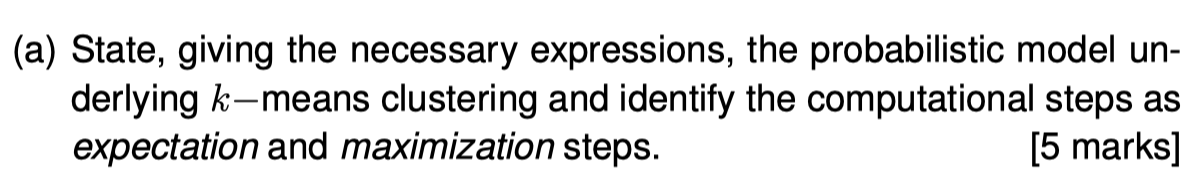
（iv）对J求偏导，取极大值。

图片包含 文字, 白板

描述已自动生成

图片包含 屏幕截图

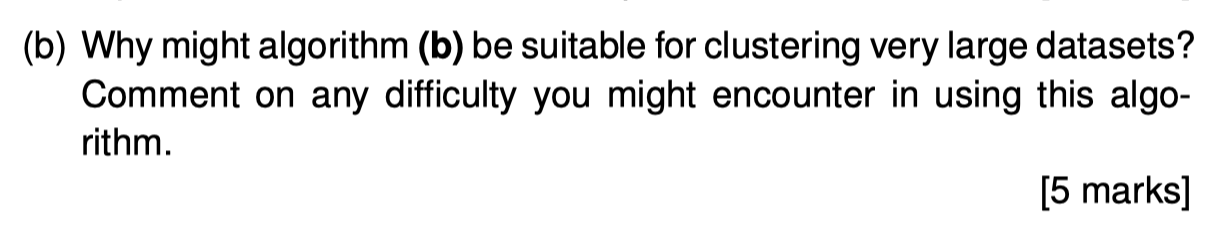
描述已自动生成



我对这道题的理解是，因为kmeans其实是高斯混合模型EM算法的特殊情况(kmeans是硬分类hard assignment，高斯混合模型是软分类soft assignment)，所以我们需要用EM的视角看待kmeans求解的过程。整个思路其实在老师的PPT中有所体现。考试中遇到就根据PPT中的内容写就可以了。当然，也可以参考这篇博客[Reference](https://www.cnblogs.com/chenjieyouge/p/12078473.html)去理解kmeans与EM。

图片包含 屏幕截图

描述已自动生成



首先先需要理解（b）算法在干嘛。遇到这种题型，先画图，这样便于理解。

1. 首先，我们先从n个数据中取出第一个数据，将其归为聚类中心。
2. 然后取出第二个数据，如果第二个数据与第一个数据距离大于theta，那么第二个数据就不属于第一个类，而是自立门户，重新成为一个聚类的中心。如图所示，我们可以看见，绿色三角因为和x数据的距离大于theta，所以它就属于一个新的聚类。
3. 从第三个数据开始，就有点不一样了。第三个数据会先判断它属于第一个数据的类，还是第二个数据的类，然后取一个距离最短的聚类中心。但是这个聚类的中心是暂时的。和第二步一样，我们还需要进行判断图片包含 物体, 手表, 时钟

   描述已自动生成。如果第三个数据和这个暂时的聚类中心距离依然大于theta，那么第三个数据也要离家出走，成为一个新的聚类中心。以此类推。
4. 刚刚讨论了如果两点之间的距离大于theta的情况。那么如果小于theta会怎么样？小于theta就要更新聚类中心点c：。这个式子就是聚类中心会向新的样本点一点一点靠近，而有点像学习率的意思。
5. 其实这道题理解还有点问题。这个是否有必要。

图片包含 文字

描述已自动生成

**为什么（b）更适合大数据集的聚类：**根据聚类中心的式子，我们可以发现，b算法在更新聚类中心的时候，是不需要遍历该类中所有的样本点的。回想一下，在Kmeans中，如果要更新当前的聚类中心，是不是要把该类下的所有点都遍历一遍然后除以N。b算法的好处就是时间复杂度低，速度快，开销低，这就适合大的数据集。

**那么b算法有什么缺点：**我能想到的缺点就是，图片包含 物体, 手表, 时钟

描述已自动生成 这两个参数如何确定下来？根据不同的数据集，可能会有不同的图片包含 物体, 手表, 时钟

描述已自动生成。也就是此算法没有很强的鲁棒性（robustness）。使用这个算法，调参(tuning parameters)很重要。但是对于kmeans来说，只需要给定一个k值，一切都可以直接搞定。

图片包含 室内

描述已自动生成

1. **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)

1. **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.

1. **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)

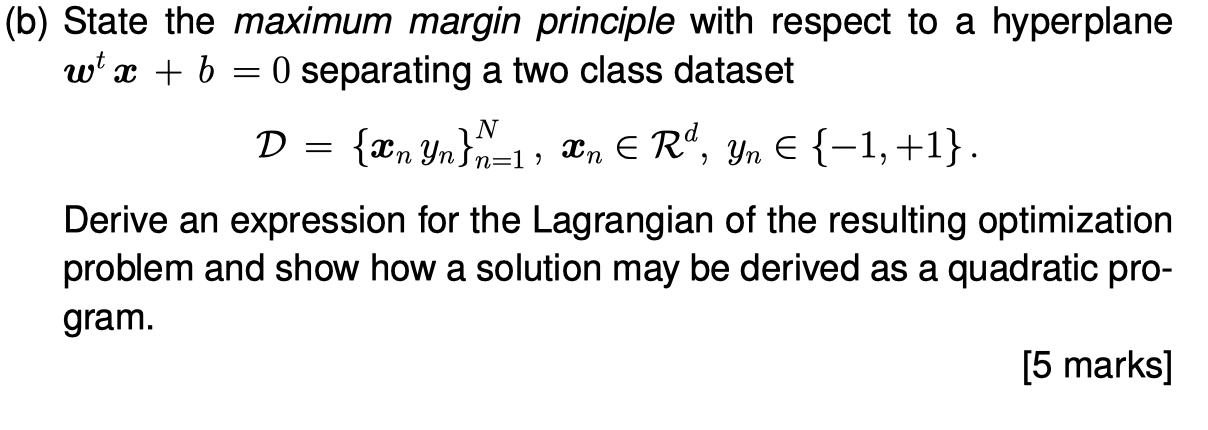
1. **提及两种避免过拟合的方法。学到的两种正则化方法。**

L1 regularization 图片包含 物体

描述已自动生成

L2 regularization 图片包含 物体

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**

