**AML**

**注：长段的句子和回答复制玩中文再放个翻译进去。公式截图放进去**。

**折叠文本快捷键：cmd + option + 8(ctrl+shift+8)将当前块转换为折叠列表** ·

2018-2019已经放上去，重复的题目直接直接在重复的题目标上+1

[软边界](bookxnotepro://opennote/?nb={1434276d-f11d-4572-9720-0d8e94e9295e}&book=6840bdfde5de360cebf83a65bd715d0f&page=7&x=323&y=338&id=1)

**去年试卷与答案：**

https://github.com/RohanBungre/COMP6208-Advanced-Machine-Learning/tree/main/Exam书签：[COMP6208-Advanced-Machine-Learning/Exam at main · RohanBungre/COMP6208-Advanced-Machine-Learning](https://github.com/RohanBungre/COMP6208-Advanced-Machine-Learning/tree/main/Exam)

# 简答题

泛化性

#### 什么是先验

The prior encodes the prior belief about the distribution of the random variable before observing the data.

在观察数据之前，先验对关于随机变量分布的先验信念进行编码。

1. 什么是likelihood

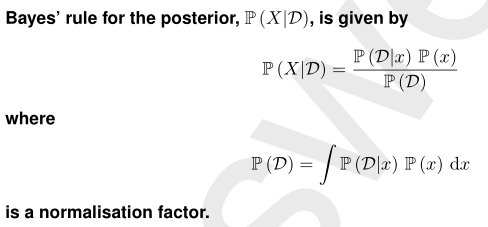
The likelihood is a model of the likelihood of the data given X

1. posterior

这个可能性是一个关于数据给定的可能性的模型

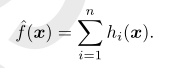
1. Bayes rule:

写出公式和标准化因子。



1. Gradient Boosting algorithm 2个问题和两个对应解决方法

梯度提升分类是一个 boosting 算法，我们通过添加弱分类hi(x)来**构建**一个**强分类器** f (x) ,



In gradient boosting we greedily build a learning machine by adding a new weak learner trained on the residues in the error from the previous layer. To control over-fitting we often use a regulariser term and additionally we use earlier stopping based on the performance on a validation set. In many implementations such as XGBoost the weak learners of regression trees.

我们贪婪地通过添加一个新的弱学习者来构建一个学习机，该弱学习者根据前一层错误中的剩余部分进行训练。为了控制拟合，我们通常使用**正则化项**，此外，我们根据**验证集的性能**使用提前停止。在许多实现中，如XGBoost，回归树的学习能力较弱。

1. Explain for **Gaussian Processes (GP)** what is the prior, the likelihood and the posterior. correlation function。distribute。

The prior is a measure over function such that the probability of points in space are **normally distributed** with a two-point correlation function given by the kernelK(x, y)—that is it is a Gaussian Process.

The likelihood is typically a Gaussian between the observed points and the **prediction** of the Gaussian process.

The posterior is also a Gaussian Process conditioned on the observations.

先验是对函数的度量，使得空间点的概率服从由核K(x，y)给出的两点相关函数的正态分布，即它是一个高斯过程。

先验是对函数的度量measure ，使得空间点的概率服从由核K(x，y)给出的两点相关函数的正态分布，即它是一个高斯过程。该似然通常是观测点和高斯过程的预测之间的高斯。

后验也是以观测值为条件的高斯过程。

1. 解释核函数代表什么，以及如何从许多观察结果中经验性地测量它。2个相似

核函数给出了两点相关函数。如果我们有足够的数据点(xi，yi)，我们就可以计算出成对距离d=||xi−xj||。然后，我们可以计算距离在某个窄区间内的所有点对的平均相关系数(yi−µ)(yi−µ)/σ²。请注意，这里的µ和σ²是整套目标值的经验平均值和方差。

**作为距离函数的函数的这种相关性应该类似于核函数。**

**此外，在每个区间内，数据应该大致呈正态分布。**

The kernel give the two point correlation function. If we have enough data point(xi, yi)we can compute the **pairwise distances** =||xi−xj||. We can then compute the **mean correlations**(yi− µ)(yi− µ)/σ²for all pairs of points whose distance lies within **some narrow interval.** Note here µ and σ² are the empirical means and variances of the **complete set of target values yi.** This correlation as a function of the distance function should be similar to the kernel function. Furthermore, within each interval the data should be roughly normally distributed.

两点相关函数—足够的数据点计算**pairwise distancesd**— narrow interval里计算每一对的平均相关。

这里的µ和σ²是整套目标值的经验平均值和方差。

作为距离函数的函数的这种相关性应该类似于核函数。

此外，在每个区间内，**数据应该大致呈正态分布。**

1. Explain why it is important that a kernel is positive semi-definite and give  
   three properties that a positive semi-definite kernel should have.

(i) Kernel functions need to be positive semi-definite so that they have sensible (non-negative) distances. That is the margins are positive.

(2 points)

(ii) The eigenvalues of a positive semi-definite kernel function are non-negative

(iii) A positive semi-definite kernel function can always be written as

K(x, y) =X i φi(x) φi(y) for some set of real functions φi(x)

(iv) The quadratic form satisfies

Zf(x) K(x, y) f(y) dx dy ≥ 0 for any real function f(x).

(i)核函数必须是正半定的，这样它们就有**可感(非负)距离**。那就是边缘是正数。

(2分)

(ii)半正定核函数的特征值是非负的

(iii)一个半正定的核函数可以写成

实函数集 φi (x)的 k (x，y) = xi φi (x) φi (y)

(iv)二次型满足

对于任意实函数 f (x) ，Zf (x，y) f (y) dx dy ≥0。

1. 解释使用MAP解决方案而不是完整的贝叶斯解决方案的优缺点。
   1. The MAP solution is usually much simpler to calculate, particularly when there is no closed form solution for the posterior. Also we don’t need to evaluate the normalisation term which often can only be computed approximately with much effort. On the other hand, if the posterior is not unimodal and strongly concentrated around its maximum value then the MAP solution can be very inaccurate. Note also that the MAP solution doesn’t give a probabilistic answer so we cannot, for example, compute the accuracy of the prediction (something you can do in the full Baysian framewwork).
   2. MAP 解通常**计算**起来要**简单**得多，特别是在**后验没有闭合解**的情况下。此外，我们不需要评估的正常化术语往往只能计算大约与很多的努力。另一方面，如果后验不是**单峰**的，强烈集中在它的最大值附近，那么 MAP 解决方案可能是非常不准确的。还要注意，MAP 解决方案**没有给出概率性的答案**，因此我们无法**计算预测的准确性**(在完整的 Baysian 框架中可以做到这一点)。
   3. 优点：对于后验close form solution ,计算更简单
   4. 缺点：如果存在local optimal solution 并且靠近global optimal solution，那么会很可能不准确
   5. 结果不是概率答案，所以无法预测结果的accuracy
2. 解释为什么蒙特卡罗技术MCMC经常被用来解决贝叶斯推理问题。—兴趣的数量

For many Bayesian problems the posterior has no closed form solution so **it cannot be expressed**. Monte Carlo techniques allow us to draw samples from the posterior which allow us to compute many quantities of interest such as the posterior mean (often the best single prediction) and the posterior variance (giving an indication of the expected error).

对于许多贝叶斯问题，**后验没有封闭形式的解**，所以它不能表示。蒙特卡罗技术允许我们从后验中提取样本，这样我们就可以计算出许多感兴趣的数量，**例如后验平均值(通常是最好的单一预测)和后验方差(给出预期误差的指示)。**

1. 用文字简要描述了MCMC算法在贝叶斯推理中的应用。—收敛到后验概率

在 MCMC 中，我们通过在满足细节平衡的可能解中做小跳跃来探索后验分布。也就是说，采取行动的可能性取决于行动前后解决方案的后验概率比率。这样可以确保采样点的分布收敛到后验概率。我们需要丢弃初始点(烧毁阶段)。这给了我们来自后部的样本点(尽管为了独立，我们需要等待这些点去相关)。

In MCMC we explore the posterior distributions by making small jumps in the possible solutions with a probability that satisfies **detail balance**. That is, the probability of making a move is dependent on the **ratio of the posterior probability of the solution** before and after the move. This ensures that the distributions of sampled points converges to the posterior distribution. We need to throw away the **initial points** (burn-in phase). This gives us sample points from the posterior (although to be independent we need to wait for the points to decorrelate).

使用满足detailed balance的tranfer probability，在马尔科夫链中 从 初始分布做一些小跳跃试图靠近后验分布，**是否才去转移**取决于转移前后解决方案的**后验概率比值**。这样采样点的分布会**converge到后验分布**。在**丢弃初始点**（一些后），就会得到满足后验分布的样本点从而计算出后验均值和后验方差

1. 什么时候概率方法可能产生好的结果，使用它的障碍是什么？

Probabilistic methods are optimal when we have an accurate **model of the likelihood of the data** and **a good posterior**. This is typically the case when we have a good understanding of how the data is generated. Probabilistic methods are expensive to use as we have to carefully model the likelihood and prior. Often we have little understanding of what generated the data, although we may have some expectation about the solution (e.g. it is likely to be continuous and not rapidly changing although these are often hard to specify as a prior). In such cases probabilistic methods are often dominated by other (more generic) machine learning techniques.

当我们对数据有对似然的准确模型和一个好的后验时概率方法是最优的。当我们很好地理解了数据是如何生成的时候，通常就会出现这种情况。概率方法使用起来很昂贵，因为我们必须仔细地建立可能性和先验的模型。**通常我们对数据的产生并不了解**，尽管我们可能对解决方案有一些期望(例如，它可能是连续的，不会迅速变化，尽管这些通常很难作为优先事项加以说明)。在这种情况下，概率方法通常被其他(更通用的)机器学习技术所支配。

当我们对数据有对似然的准确模型和一个好的后验时概率方法是最优的，这代表我们能学习到数据的结构。or 是如何生成的时候，通常就会出现这种情况

缺点：概率方法的使用**代价很高**expensive to use 因为需要建立准确的**后验模型和先验模型。**

缺点2：通常情况下我们**很难对数据的结构有足够的了解**。这些情况下概率模型的表现往往低于其他机器学习方法。

1. Explain the bias and the variance terms in words.

1。偏差是平均值机器的泛化性能(即，通过使用给定大小的每个可能的训练集训练的机器的平均响应来进行预测)。

2。方差衡量各台机器的反应如何因机器而异。

(i) The bias is the generalisation performance of the mean machine (i.e. the prediction made by taking the mean response averaged over machines trained with every possible training set of a given size).

(ii) The variance measures how the response of the individual machines vary from machine to machine.

1. What is the bias-variance dilemma.

To get good generalisation we want to both reduce the bias and variance. To reduce the bias we need a complex machine, but that will typically to increase the variance. Conversely a simple machine is likely to have a small variance, but a high bias.

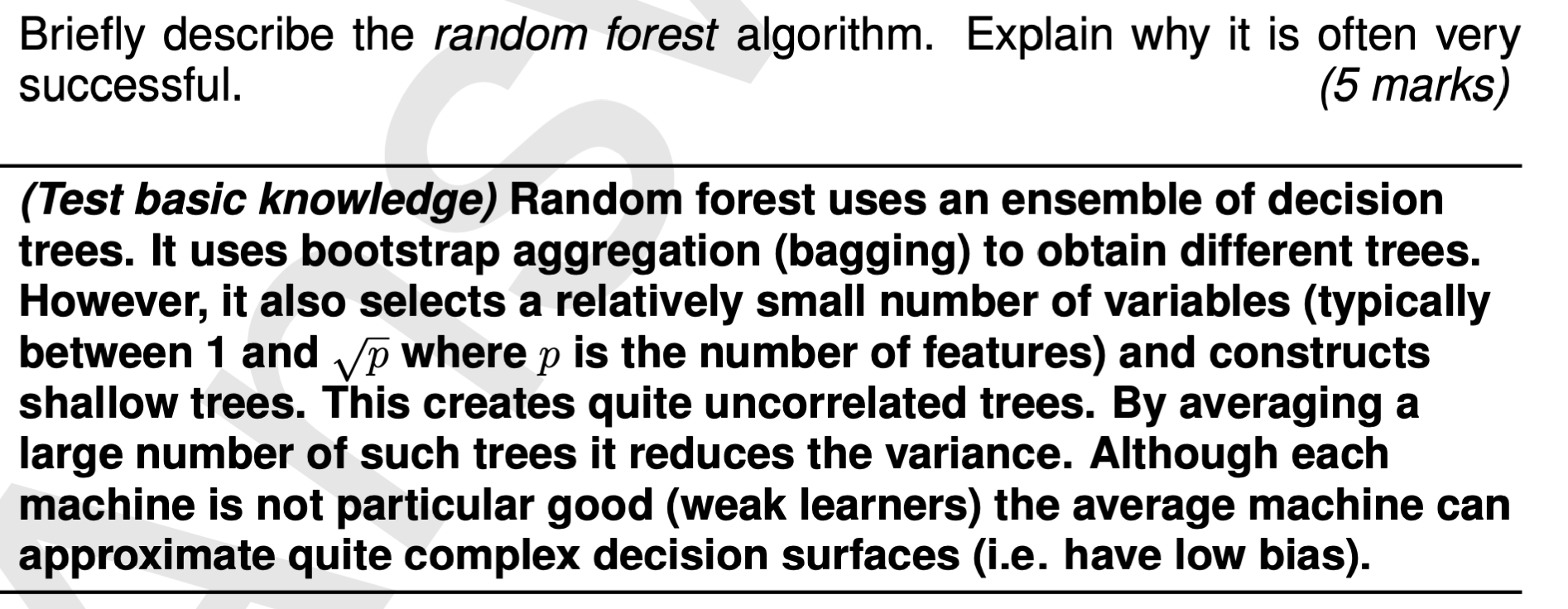
为了得到良好的一般化，我们要减少偏差和方差。为了减少偏差，我们需要一个复杂的机器，但这通常会增加方差。相反，一台简单的机器很可能有小的方差，但偏差很大。

1. 解释随机森林算法如何尝试克服偏差-方差困境。decision trees

Random forest **averages** a **large number** of decision trees that **have been trained** on **slightly different data sets** (through bootstrapping) using a different set of features. This averaging reduces the **variance**.

随机森林平均大量的决策树，这些决策树已经**在稍微不同的数据集上训练过**(通过引导)使用不同的特征集。这个平均数减少了方差。

1. (c) 简要描述随机森林算法。解释为什么它经常⾮常成功。4点(5 分)



（测试基础知识） 随机森林使用**决策树的集合ensemble**。它使用引导聚合（**装袋**）来获得**不同的树**。

然而，它也选择了相对较**少数量的变量**（通常在 1 到 ✓p 之间，其中 p 是特征的数量）**并构建浅树**shallow。

这会创建**完全不相关的树**。

通过对大量此类树进行平均，可以**减少方差**。

尽管每台机器都不是特别好（弱学习器），但平均机器可以**逼近相当复杂的决策面**（即具有低偏差）。

**决策树的集合ensemble**—bagging **在稍微不同的数据集上训练过**

1. 解释正规化如何在偏差-方差困境的背景下起到帮助作用。

By introducing a regularisation term we can use powerful machines (with potentially low bias), however, the regulariser usually makes the learning machine less sensitive to the data, thus reducing the variance.

通过引入正则化项,**在稍微不同的数据集上训练过**从而减少方差，因此我们可以使**用强大的机器**(具有潜在的**低偏差**)，

1. bagging

Bagging 是一种集成学习技术，可以在许多不同的机器上进行平均。 每台机器都在不同的数据集上进行训练。 数据集是通过**bootstrapping**创建的。 也就是说，我们对训练集进行替换以创建新的训练集。 我们在每个数据集上训练一个学习机，然后取这组机器的平均响应。

这通过对不同数据集进行平均来**减少方差**（在偏差-方差困境中），从而提高泛化性能。

使用 bagging 的著名机器学习算法是对决策树进行平均的随机森林算法。

1. begging is

## SVM

1. Describe what you need to do practically to ensure that SVMs work  
   well.

通常重要的是**确保特征被归一化**(例如，通过减去平均值并除以标准差)。

optimise over the punishment term,C, for the **slack variables**和**核的任何参数**(例如，径向基函数核中的γ变量)进行优化也很重要。

这通常涉及对许多数量级执行网格搜索-当从网格的一点移动到邻居时，通常会将参数增加一倍。

It is usually important to ensure that the features are normalised (e.g. by subtracting the mean and dividing by the **standard deviation**). It is also important to optimise over the punishment term,C, for the slack variables and over any parameters of the kernel (e.g. the γ variable in RBF kernels). Typically this involves performing a grid search over many orders of magnitude—often doubling the parameters when moving from one point of the grid to a neighbour.

1. (a)Explain why choosing the maximum margin dividing plane is so important to the success of SVMs . ( 5 marks )当使用扩展的特征空间时 ， 这是必不可少的 ， 否则会**具有巨大的过度拟合数据集的能力 。**

By choosing the maximum margin hyperplane the SVM finds a solution that tends to generalise well . That is , it is the dividing plane that can **tolerate** the largest level of noise on the data points and still give the correct classification . This v is essential when using an extended feature space which would otherwise have a huge capacity to over - fit the data set .  
翻译：

( 容易开始 ， 但理解很重要 。 )

通过选择最大边距超平面 ， SVM 找到了一个易于泛化的解决方案 。 也就是说 ， 它是能够容忍数据点上最大程度的**噪声**并且仍然给出正确分类的分割平面 。 当使用扩展的特征空间时 ， 这是必不可少的 ， 否则会**具有巨大的过度拟合数据集的能力 。**

1. kernel
2. 综合
   1. 判别模型和生成模型
      1. 判别概率模型预测目标 y 给定特征 x 的概率，即 P (y | x)。 在生成模型中，我们使用联合概率 P (y, x) **生成目标和特征**。 **判别模型通常更容易，因为我们不需要对生成特征的过程进行建模**。 生成模型可以提供更准确的模型，并且可以以许多不同的方式使用。
   2. svm和gb的适用范围
      1. (i) SVM 擅⻓处理⼩型⾼维数据集（平衡良好，⽆缺失数据）  
         (ii) 梯度提升适⽤于混乱的表格数据（缺失数据、混合类型等）
3. w距离
   1. KL不是真正的距离而w是
   2. 两个分布温泉没有重叠，KL的值是没有意义的
   3. w距离：**即便两个分布没有重叠，Wasserstein距离仍然能够反映它们的远近**
   4. KL散度和JS散度是突变的，要么最大要么最小，**Wasserstein距离却是平滑的**Smooth
4. 最大熵

### 最大熵法

* + 当我们试图**推断给定观测值的分布**时，我们可以在约束条件下**最大化熵-**熵作为一个先验
  + 这就是所谓的最大熵方法
  + 我们可以对此进行合理化解释，因为这是迄今为止最有可能与观测结果一致的一组配置
  + 或者，我们可以把这看作是我们已知情况下**最大化不确定性的**做法
  + 只有在所有概率相等的情况下，它才能给出一个好的近似值。**对取值离散的随机过程，最大熵意味着取平均概率，也就是说当取值概率相等的时候，熵最大化**
  + 从上面可以看到，在给**定约束条件下，基于最大熵原理可以得到某种概率分布函数**，给定常量的均值和方差，可以得出正态分布，这个思路指明了不同的约束会导致不同的概率分布结果。

1. SGD Explain what is meant by i) Stochastic Gradient Descent ii) momentum in the context of learning and iii) mini-batches.

(i) Stochastic gradient descent means that we compute the gradient and update the weights after each image (or mini-batch)—2 points

(ii) Momentum is when our update “velocity” remembers previous updates that are slowly changed by the current gradient—2 points

(iii) Mini-batches is when we compute the average gradient for a small number of input patterns (e.g. 5–100) at each time step and update according to this average gradient—1 point.

(i)随机梯度下降是指我们计算梯度和更新权重后，每个图像(或小批量)ー2点

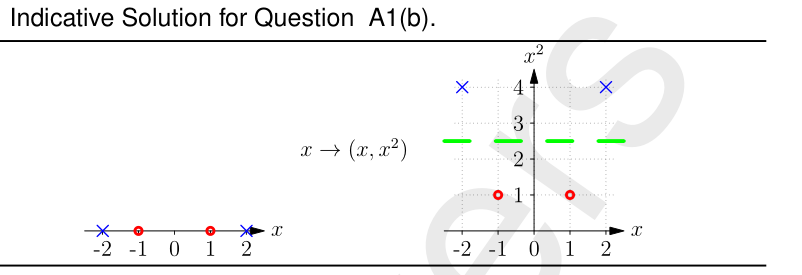
(ii)动量是当我们的更新“速度”记住以前的更新，这些更新被当前的梯度慢慢改变ー2点

(iii)小批量是指我们在每个时间步骤计算少数输入模式(例如5-100)的平均梯度，然后根据这个平均梯度ー1点更新。

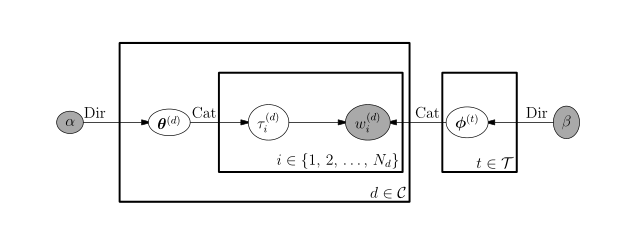
## graphical model

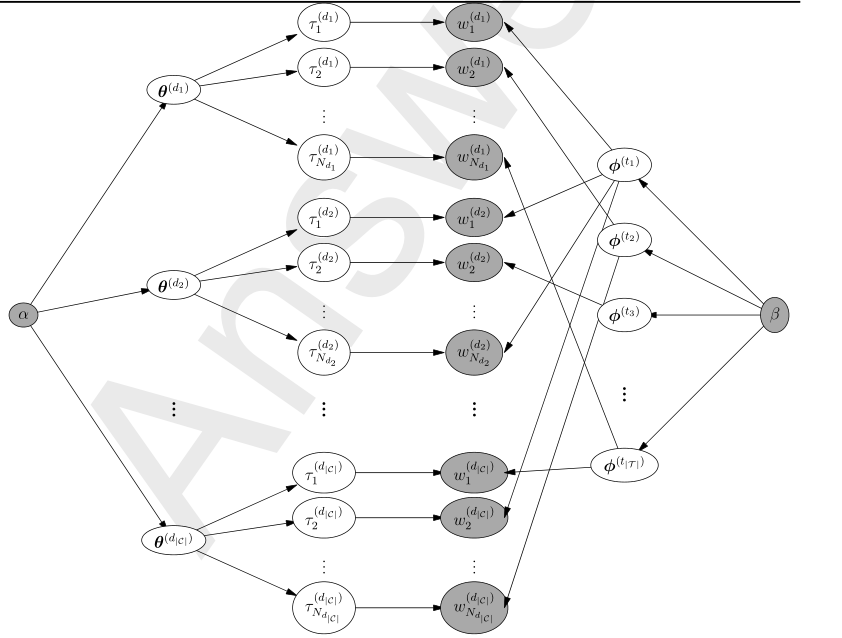
# 画图题

SVM

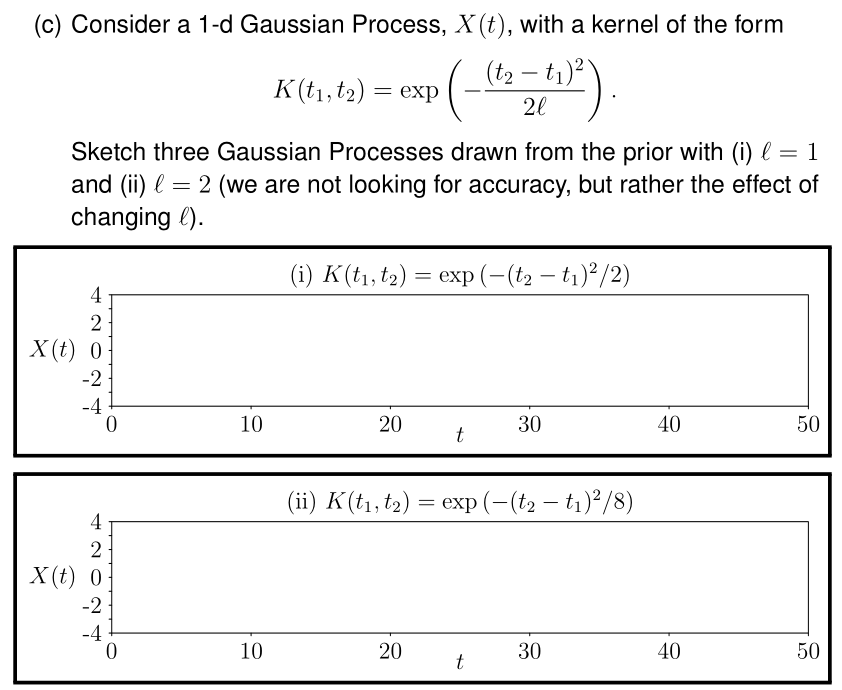


1. show the full word generation process.—graphical model



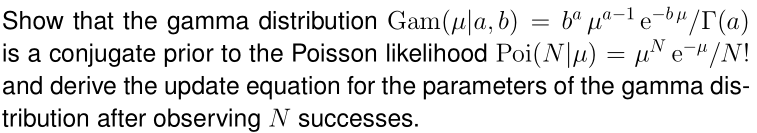


1. GP

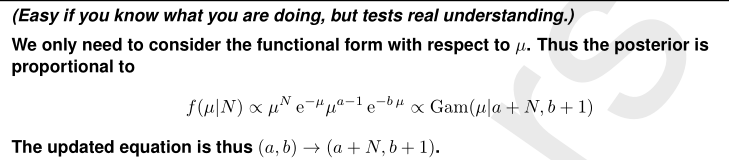


# 推导题一

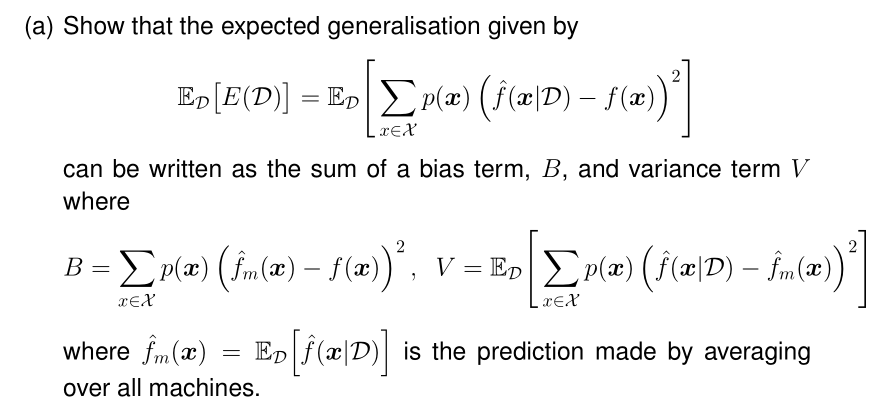
1. 共轭先验

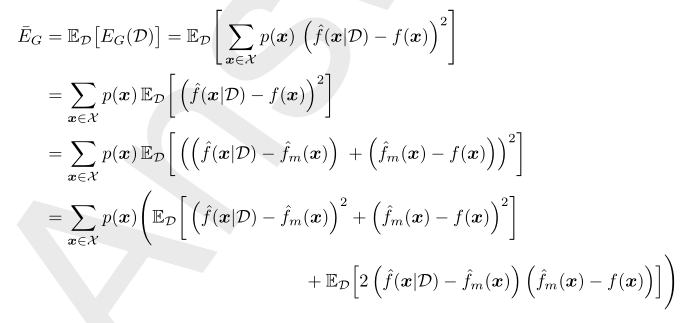


我们只需要考虑函数形式对 μ 的影响，因此后验函数与 μ 成正比



1. 用bias和variance来表示泛化性





**上面的内容按主题分类用嵌入放进下面**：复制方法：在上面选中块，点击Q。即可复制块引用，然后ctrl+v粘贴到下面

# 重点主题：

注：里面每个主题的建议了解是什么，表现好的条件是什么，有什么用，优点，缺点。需要大家一起补全一下。

试卷里的**出现过的主题用黑体加粗**，ppt出现试卷没出现过的不加粗

**如果试卷里出现没加粗的主题，就去加粗一下**

## SVM

## 概率模型

### MCMC

蒙特卡洛方法

拒绝方法采样

### GP

1. Explain for **Gaussian Processes (GP)** what is the prior, the likelihood and the posterior.

(Conceptually challenging. This question is from last year, but almost no one tackled it.) The prior is a measure over function such that the probability of points in space are normally distributed with a two-point correlation function given by the kernelK(x, y)—that is it is a Gaussian Process.

The likelihood is typically a Gaussian between the observed points and the prediction of the Gaussian process.

The posterior is also a Gaussian Process conditioned on the observations.

先验是对函数的度量measure ，使得空间点的概率服从由核K(x，y)给出的两点相关函数的正态分布，即它是一个高斯过程。该似然通常是观测点和高斯过程的预测之间的高斯。

后验也是以观测值为条件的高斯过程。

### MAP和MLP

**Gradient boosting**

**共轭先验**

1. 矩阵相关知识

试卷没出现过，但ppt中间经常提及

* 1. 奇异矩阵

正交矩阵，

Cross Validation

em算法

**随机森林**

**GB**

## PPT内容：

总结一下PPT中出现的算法or方法名称，用于和重点主题对照，看看那些主题发生变化

### bias and variance

bias, variance 以及泛化误差

泛化：

提高泛化性能  
把你的数据分成训练和验证集  
不能使用你训练过的数据  
使用验证集选择超参数  
真正聪明的机器会尝试自动完成这种匹配

### overfiting

Cross Validation

### projects

### ensembleLearning

### vectorSpace

### mapping

### PCA

### SVD

### optimisation

### SGD

### 约束优化

### 凸函数

### svm

### 核技巧

### wasserstein距离

### 贝叶斯

### GP 高斯过程

### 概率推理

em算法

### 图模型

### MCMC

### 熵

### 信息论

解释正规化如何在偏差-方差困境的背景下起到帮助作用。

解释正规化如何在偏差-方注释1差困境的背景1注释2下起到帮助作用。

注释3

[注释1] 特侧

[1] 差-方差困境的背景

[注释2] 123

[注释3] 123