2017-2018-1 CS28010 类脑智能期中考试

隋国新

2017 年 11 月 27 日

1 （视觉理论）请用自己的语言解释

(a) Wiesel-Hubel 特征检测理论

(b) Marr 的视觉计算理论

(a) Hubel 和 Wiesel 等通过对猫的大脑视觉皮层系统的研究，发现了视觉皮层通路中特定的神经元对特定方向的直线敏感，还有一些神经元对复杂的模式敏感，进而推断出视觉皮层先发现简单信息，然后组合为复杂模式的信息的分层处理机制。

(b) Marr 从信息处理系统的角度出发，认为视觉系统的研究应分为三个层次，即计算理论层次、表达与算法层次、硬件实现层次
计算理论层次要回答系统各个部分的计算目的与计算策略，亦即各部分的输入输出是什么，之间的关系是什么变换或者具有何种约束。Marr 对视觉系统的总的输入输出关系规定了一个总的目标，即输入二维图像，输出是由二维图像” 重建”（reconstruction）出来的三维物体的位置与形状。
对于表达与算法层次，视觉系统的研究应给出各部分（模块）的输入、输出和内部的信息表达，以及实现计算理论所规定的目标的算法。
硬件层次: 是要回答” 如何用硬件实现以上的算法”。

2 （回归分析）感知机 perceptron

(a) 用自己的语言结合公式解释什么是 perceptron

(b) 请解释为什么用 perceptron 没有办法解决 XOR（异或）的求解

(c) 如何变化 perceptron 使之解决问题 b)?

(a) MP 神经元模型中，神经元接收到来自 n 个其他神经元传递过来的输入信号，这些输入信号通过带权重的连接进行传递，神经元接收到的总输入值将与神经元的阈值比较，通过激活函数处理产生神经元的输出。
perceptron 由两层神经元组成，输入层接收外界输入信号后传递给输出层，输出层是一个 MP 神经元。

(b) XOR 是一个非线性可分问题。perceptron 只有一层功能神经元，而 MP 神经元本质上是对输入进行线性处理后通过一个单调的激活函数，故不能解决这样的问题。

(c) 可以使用多层功能神经元。

3 （最小平方拟合）数据集 $x_i, i = 1, ..., n$，其中 x_i 是一个 d 维向量。假设它们的均值为 0。现有一个单位长向量 w。x_i 沿着向量 a_i 在 w 上的投影为 $\hat{x}_i$

(a) 请写出 $\hat{x}_i$ 关于 x_i, a_i 和 w 的表达式

(b) 求出使 $\Sigma_{i=1}^{n}||\hat{x} - x_i||^2$ 最小的 a_i 和 w

(a) 由题意知 $\hat{x}_i$ 的方向为 w，可以设 $\hat{x}_i = kw$。设 $a_i = \hat{x}_i - x_i$，即 $\hat{x}_i, x_i, a_i$ 恰好是一个三角形的三条边，则 $\hat{x}_i = a_i + x_i$

(b) 令 $J = \sum_{i=1}^{n} ||\hat{x}_i - x_i||^2 = \sum_{i=1}^{n}(k_i w_i - x_i)^T(k_i w_i - x_i)$
令 $\frac{\partial J}{\partial k_i} = 0$，得到 $k_i = w^T x_i$，可知 $\hat{x}_i$ 是 x_i 的正交投影。
因为 w 是单位向量，故满足 $w^T w = 1$，
所以可以构造拉格朗日算子 $L = \sum_{i=1}^{n}(k_i w_i - x_i)^T(k_i w_i - x_i) - \lambda(w^T w - 1)$ 令 $\frac{\partial L}{\partial w} = 0$，得到 $\Sigma w = \lambda w$，这里 Σ 是 X 的协方差矩阵。
可知 w 为特征向量，λ 为特征值，w 取最大的特征值对应的特征向量的时候 $\Sigma_{i=1}^{n}||\hat{x} - x_i||^2$ 最小。

4 （高斯分布）$G(x|\mu_x, \sigma^2 I) = \frac{1}{(2\pi)^{d/2}\sigma d} exp(-\frac{1}{2\sigma^2}(x - \mu_x)^T(x - \mu_x)$，其中 x 为 d 维向量。$G(y|\mu_y, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} exp(-\frac{1}{2}(x - \mu_y)^T\Sigma^{-1}(x - \mu_y)$，其中 y 为 d 维向量。X 和 y 之间满足 $x = By$

(a) 请写出 B 的表达式

(b) 请针对 B 表达式的形式来解释 B 的含义 (对 y 做了什么操作得到 x)

(a) 根据题意我们有：

$$\begin{aligned} E(x) &= E(By) \\ &= BE(y) \\ Cov(x) &= Cov(By) \\ &= E[(By - E(By))(By - E(By))^T] \\ &= E(Byy^T B^T - BE(y)y^T B^T - ByE(y)^T B^T + BE(y)E(y)^T B^T) \\ &= E(Byy^T B^T - B\mu_y \mu_y^T B^T) \\ &= B(E(yy^T) - \mu_y \mu_y^T)B^T \\ &= B\Sigma B^T \end{aligned}$$

即：$\begin{cases} \mu_x = B\mu_y \\ \sigma^2 I = B\Sigma B^T \end{cases}$

(b) 由 $\mu_x = B\mu_y$ 可知将 y 的均值移动至 x 的均值处，由 $\sigma^2 I = B\Sigma B^T$ 可知将 y 的方差压缩，对角线方向调整至 σ，反对角线方向分布密度与对角线方向一致。

5 （统计决策）请用自己的语言结合公式解释

(a) Bayesian classification

(b) Fisher discriminant analysis

(c) 它们的区别

(a) 根据贝叶斯定理，$P(c|x) = \frac{P(c)P(x|c)}{P(x)}$，其中 $P(c)$ 是分类为 c 的先验概率，$P(x|c)$ 是给定类标记后样本 x 的条件概率。Bayesian classification 就是根据这个公式，基于训练数据估计 $P(c)$ 和 $P(x|c)$，进而求出后验概率 $P(c|x)$。

(b) 费舍尔两类分布的规则为类间方差与类内方差的比率 $S = \frac{\sigma^2_{between}}{\sigma^2_{within}} = \frac{(w\mu_1 - z\mu_0)^2}{w^T\Sigma_1 w + w^T\Sigma_0 w}$，其中 $\mu_0, \mu_1, \Sigma_0, \Sigma_1$ 分别为两个中心的均值和方差，wx 是特征的线性组合。
Fisher discriminant analysis 就是根据这个公式，进行投影，将原来一个维度空间的自变量组合投影到另一维度空间，寻找一个由原始变量组成的线性函数使得组间差异和组内差异的比值最大化。

(c) 1. 原理不同
Bayesian classification 是是利用已知的先验概率去推证将要发生的后验概率，就是计算每个样本的后验概率及其判错率，用最大后验概率来划分样本的分类并使得期望损失达到最小
Fisher discriminant analysis 是根据方差分析思想，利用投影使组间差异和组内差异的比值最大化。根据样本点计算判别函数，计算判别函数到各类中心的欧式距离，取距离最小的类别。
2. 前提条件不同:
Fisher discriminant analysis 不考虑样本的具体分布，只求组间差异和组内差异的比值最大化
Bayesian classification 从样本的多元分布出发，充分利用多元正态分布的概率密度提供的信息计算后验概率，因此需要样本数据服从多元正态分布，方差齐性等。

6 （高斯因子分析）请用自己的语言结合公式解释 linear Gaussian factor analysis($X = Ay + e$) 的

(a) scale 不确定性

(b) rotation 不确定性

(c) additive 不确定性

(d) dimension 不确定性

(e) 若 factor y 的分布从高斯分布变成二项分布, 上述不确定性还存在么? 为什么?

定义 A 是一个 $n \times m$ 的矩阵, n 是 x 的维数, m 是 y 的维数，$m < n$

(a) scale 不确定性: y 服从于一个 m 维的高斯分布，均值与方差的 scale 都不确定，并且可以随着 A 的 scale 做出相应调整，存在 scale 不确定性

(b) rotation 不确定性：若不对 y 的方差做特殊限制，那么 y 可以旋转，存在 rotation 不确定性

(c) additive 不确定性: 考虑 Ay 与 e 的值的分配，二者的尺度都不确定时，additive 不确定性存在

(d) dimension 不确定性：y 的维度 m 不确定，存在 dimension 不确定性

(e) 若 factor y 的分布从高斯分布变成二项分布,scale 不确定性，rotation 不确定性，消失只有 dimension 不确定性和 additive 不确定性。

7 （最小平方聚类）

(a) 用自己的语言结合公式解释什么是最小平方聚类

(b) 就最小平方聚类而言, 在计算条件允许的情况下, 给定类别数 k, 如何求最优聚类?

(c) k-means 有三个无法解决的问题, 它们分别是什么? 为什么会存在?

(d) 但是因为代价太高, 一般会使用近似算法来求局部最优解。请用 EM 的思路来解释这个近似求解算法。(需要公式推导过程)

(a) 给定样本集 $x_i, i = 1, ..., m$, 聚类后获得簇划分 $C_i, i = 1, ..., k$，最小平方聚类方法最小化平方误差 $E = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||_2^2$，E 值越小，簇内样本相似度越高。

(b) 每一个数据点的分类有 k 种，所有点的分类有 k^m 种，计算所有情况的平方误差，取最优解

(c)
- 必须事先给出 K，而且对初值敏感，对于不同的初始值，可能会求出不同的局部最优解
- 只能发现球状簇，不适合于发现非凸形状的簇或者大小差别很大的簇
- 对噪声和孤立点数据敏感，如簇中含有异常点，将导致均值偏离严重。

(d) 假定样本以各自的簇为中心满足高斯分布，则可以使用 EM 混合高斯模型。根据 EM 算法，

- 选择初始的 K 个类别中心 $\mu_i, i = 1, ..., k$
- E-step: $Q_i(c^i) = p(c_i|x_i, \theta)$
  将软分类退化为硬分类，则有 $c_i := argmax(p(c_i|x_i; \theta))$
  不考虑高斯分布方差，则有 $c_i := argmin||x_i - \mu_i||^2$
  即对于每个样本 x_i，将其标记为距离类别中心最近的类别
- M-step: $\theta := argmax \sum_i \sum_{c^i} log \frac{p(x_i, c_i; \theta)}{Q_i(c^i)}$
  根据上步结果，更新簇中心为 $\mu_j := argmax \sum_i logp(x_i, c_i; \theta) = \frac{\sum_{i=1}^{m} c_i = j x_i}{\sum_{i=1}^{m} c_i = j}$
- 重复前两步，直到类别中心的变化小于某阈值或者达到最大迭代次数

8 （内向和外向理论）

(a) 请用自己的语言解释 best abstraction(最佳抽象) 和 best reconstruction(最佳重建) 理论

(b) 请使用最佳抽象和最佳重建理论从两个不同角度解释主成分分析 (principal component analysis)(需要公式推导过程)

(a) 最佳抽象：原始数据高维无序，剔除无关特征保留主要特征实现数据降维而不丢失信息，即信息之最佳保持。
最佳重建：利用原始数据的主要特征建立误差最小的数学模型。

(b) 给定样本集 $x_i, i = 1, ..., n$, 将 x_i 投影到 w 方向上为 $\hat{x}_i$。
按照最佳抽象理论，需要最大化保持数据差异，即 $w = argmax \sum_{i=1}^{n} (x_i^T w - \mu_x^T w)^2$。
不失一般性，令 $x_i = x_i - \mu_x$，则 $w = argmax \sum_{i=1}^{n} (x_i^T w)^2 = argmax w^T \left(\sum_{i=1}^{n} (x_i^T)^2 \right) w$ w 取最大的特征值对应的特征向量的时候目标获得最大值。PCA 将 n 维的 x 变成了 k 维的 y 实现了数据降维且不丢失信息

按照最佳重建理论, 需要最小化误差，即

$$\begin{aligned} w &= argmin\Sigma_{i=1}^{n}||\hat{x_i} - x_i||^2 \\ &= argmax - \Sigma_{i=1}^{n}||\hat{x_i} - x_i||^2 \\ &= argmax\Sigma_{i=1}^{n}\left(||x_i||^2 - ||\hat{x_i} - x_i||^2\right) \\ &= argmax\Sigma_{i=1}^{n}(x_i^T w)^2 \end{aligned}$$

随后与之前 BA 方法的求解方法一致，并有相同解。PCA 用 k 维的 y 表示 n 维的 x 保留了数据的主要特征。

9 （附加题）请用自己的语言结合例子解释五行理论（A5）对日常科研学术的启发

五行理论 (A5):
对于一个通用问题的解决, 其过程往往可以用五个主要机制的循环来表达: A-1 (Acquisition) : 获取数据 A-2 (Assumption): 基于 A-1 的结果来对内在表达作出一或多个假设 A-3 (Amalgamation): 分配和更新支持这些候选的证据 A-4 (Apex-seeking): 根据这些证据决定一或多个最好的候选 A-5 (Affirmation): 通过测试样例来评估这些候选, 并作出最终的肯定这五个行为顺序循环进行, 并且遵循各自内部的规则, 具有相对独立性.
对日常科研的启发
在实验过程中，通常我们初步实现一个模型后，结果会有不同的问题。解决问题的有效方法是分析问题可能的原因，从源头去解决。
例如，我们用随机森林去实现二分类算法，按照 A5 理论，A1 过程进行得不好会直接影响 A2 的进行，A4 过程如果进行得过深就会导致常见的过拟合问题，需要 A5 来抑制 A4 过程的过分展开。这些相互依存关系正类似于五行理论中的”生克乘晦”，通过调整”克”顺序中前一个元素来使其趋向于平衡的状态。此例中，如果 A5 中的准确率过低，可能需要调整 A4 训练参数；A3 的结果不好处理，可能是 A1 过程中留下的必然因素。