Lecture 04: 模式识别与参数估计

Kai Yu and Yanmin Qian

Cross Media Language Intelligence Lab (X-LANCE)
Department of Computer Science & Engineering
Shanghai Jiao Tong University

2021

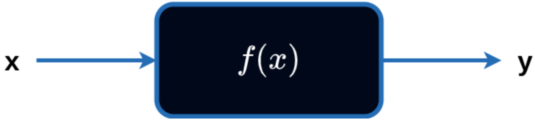
目录

回顾：多个随机变量的概率

- **联合概率**：$P(X,Y)$ 是 X 和 Y 同时发生的概率。
- **条件概率**：$P(X|Y)$ 是在 Y 的条件下 X 发生的概率。
- **边缘化**：给定 $P(X,Y)$，对 Y 的所有情况计算 X 的概率。

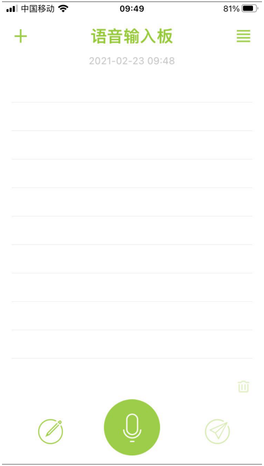
Q：后验和先验概率是什么？

机器学习和模式识别就是通过一定的数学框架，用机器来模拟人类的学习过程。



x：观察所得的输入信息。
y：期望的输出信息。通常是实数或整数（向量）

- 机器学习
  - $f(x)$ 的数学性质
- 模式识别
  - $f(x)$ 的设计与实现

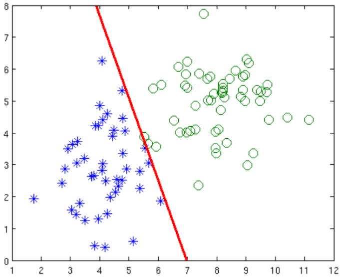


- 科学
  - 隐马尔可夫模型
  - 深度神经网络
  - 加权有限状态变换机
- 工程
  - 优化和搜索
  - 大规模计算
  - 实时实现
  - 精细归一化

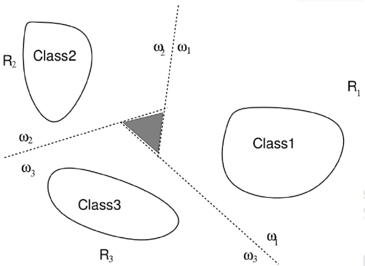
- 分类
  将每个输入分配至一组类别中的一个。
- 回归
  对每个输入指派一个实值输出。
- 序列标注
  对序列中的每个元素进行分类或回归。
- 决策
  指派一系列动作来获得最优的总体回报。

做什么？
分类

将每个输入分配至一组类别中的一个



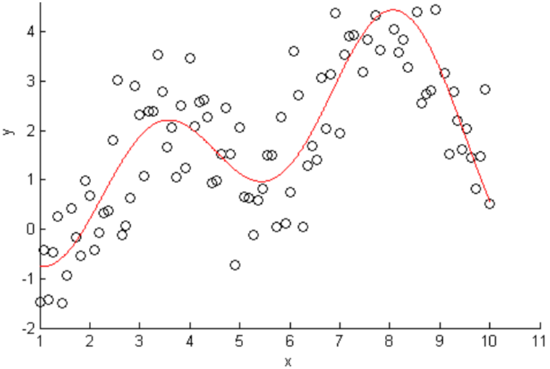
二分类



多分类

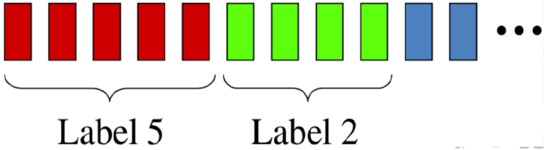
做什么？
回归

对每个输入指派一个实值输出

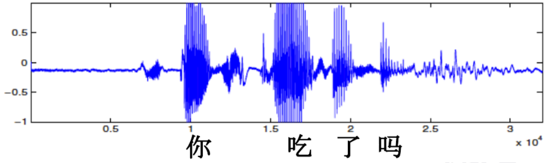


对序列中的每个元素进行分类或回归

对齐的

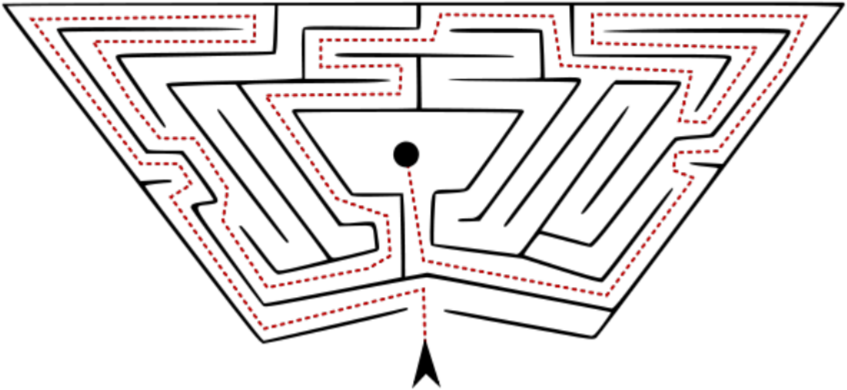


非对齐的



做什么？
决策

指派一系列动作来获得最优的总体回报



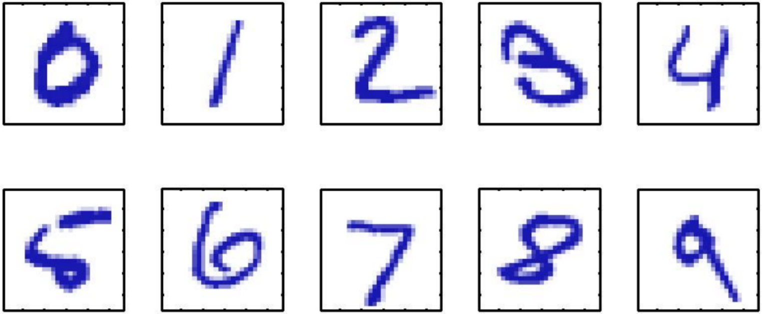
Q： 还有其它的机器学习或模式识别目标吗？

机器学习和模式识别的两个阶段

- **训练**（或学习）阶段
  - 训练数据 - 通常是（输入，输出）成对
  - 确定 $f(x)$
- **测试** 阶段
  - 测试：分类/预测/解码 ⋯
  - 泛化问题

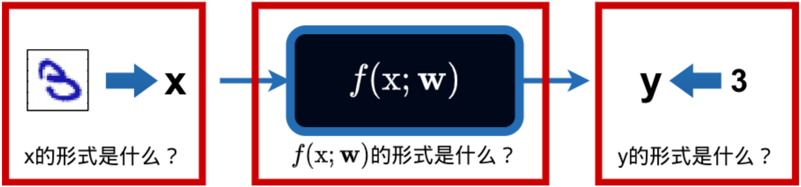
机器学习和模式识别示例

手写数字识别



手写数字识别问题

基于判别性函数的数学表述

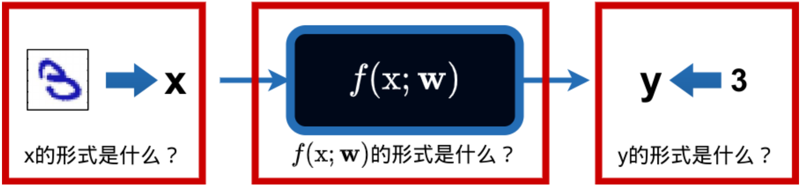


判别性识别函数

- 训练
  - $(x_i, y_i), i = 1, \ldots, N \Rightarrow \mathbf{w}$
- 测试
  - $\mathbf{w}, x_j \Rightarrow y_j$

手写数字识别问题

基于贝叶斯理论的数学表述



- 训练
  - $(x_i, y_i), i = 1, \ldots, N \Rightarrow p(x, y; \mathbf{w})$
- 测试
  - $\mathbf{w}, x_j \Rightarrow y_j = \arg\max P(y_j | x_j)$

实用机器学习的要素

- 机器学习的类别
- 数据（表示）
  - 输入（观测特征）
  - 输出（标签）
- 模型
  - 参数化或非参数化的
  - 模型的形式（参数是什么）
- 准则
  - 最优的含义是什么
- 机器学习的阶段
  - 训练
  - 测试（分类/回归/搜索···）

微博精神病患者追踪

学生期末成绩预测


说 明
学年
每学年包括秋季、春季两个学期，学校还开设暑期教学课程。每个学期包括18个教学周。理论课程每周一个课时，上满18周，获一个学分。实践、实验、听力、体育等课程由各教学单位自主安排时间和学分比。
考核与计分方式
1、考试成绩按百分制记载。分数、等级与绩点的换算关系如下表：

等级	A	A-	B+	B	B-	C+	C	C-	D	D-	P	F
绩点	4.0	3.7	3.3	3.0	2.7	2.3	2.0	1.7	1.3	1.0	0	0
百分制参考标准	90-100	85-89	82-84	78-81	75-77	71-74	66-70	62-65	60-61	补考及格	合格	59及59以下

学分绩点的计算办法是：
一门课程的学分绩＝绩点×学分数；
学期或学年的平均绩点＝所学课程学分绩之和÷所学课程学分数之和。
2、标志"*"者为外校转入成绩，计学分，不计绩点。
3、成绩单记载学生在校期间修读的所有课程成绩。根据《复旦大学学士学位授予工作细则》，授予毕业生学士学位时，以符合学生毕业时所在专业教学培养方案规定的所有课程的有效成绩为准计算平均绩点，如达到学士学位的绩点标准，并满足其他相应条件，授予学士学位。
复旦大学
FUDAN UNIVERSITY
Transcript of Academic Record
学生成绩单

机器学习类别选择

中文分词

周恩来到北京

- 周恩　来到　北京
- 周恩来　到　北京

分类问题的若干基本概念

- **两个不同的步骤**：
  - 特征提取：从原始输入中提取特征（观测）
  - 识别/推理：为观测指派一个已知类别
- **两个数据集合**：
  - **训练数据**：用于调整分类器的结构或参数
    - **有监督**：观测值和对应的正确标签均已知
    - **无监督**：仅观测值已知
  - **测试数据**：预测标签由分类器来确定
- **两类模型**：
  - 描述边界：线性分类器
  - 描述区域：概率分布
- **关键问题**：
  - 在训练数据上提高准确率
  - 泛化至测试数据和集外数据

贝叶斯公式给出了两个随机变量之间的关系

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

- **先验或边缘概率**: $P(X)$ 是 X 的**先验或边缘**分布，因为它没有考虑任何 Y 的信息
- **后验或条件概率**: $P(X|Y)$ 描述了 X 在 Y 条件下的概率，也就是说给定 Y 发生时 X 的概率

基础贝叶斯学习

- 模型
  - 古典参数化概率分布
- 训练
  - 参数估计（准则/估计）

$$\hat{\theta} = \arg\max_\theta \mathcal{L}_{\mathtt{train}}(\theta; \mathcal{D}) \quad \mathcal{D} = \{(\mathbf{x}_1, \mathbf{t}_1), \cdots, (\mathbf{x}_N, \mathbf{t}_N)\}$$

- 测试
  - 贝叶斯决策理论

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} \mathcal{L}_{\mathtt{test}}(\mathbf{x}, \mathbf{t}; \theta)$$

- 推理步骤
  - 确定 $p(t|\mathbf{x})$ 或 $p(\mathbf{x}, t)$
- 决策步骤
  - 对给定的 $\mathbf{x}$，确定最优的 t

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} \mathcal{L}_{\texttt{test}}(\mathbf{x}, \mathbf{t}; \theta)$$

- 推理和决策很多时候直接统一在一起

$$\mathcal{L}_{\texttt{test}}(\mathbf{x}, \mathbf{t}; \theta) = p(\mathbf{t}|\mathbf{x}; \theta)$$

分类的贝叶斯决策规则

总体目标：如何构建一个最小化平均错误率的模式分类器？
考虑这样一个系统：

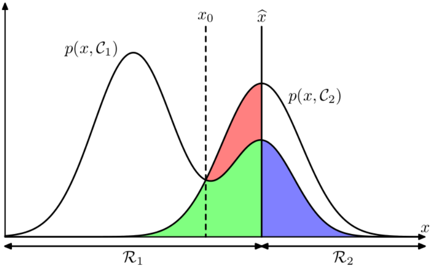
- 特征向量 $\mathbf{x}$
- K 个类别：$w_1, w_2, ..., w_K$
- K 个先验的集合：$P(w_1), P(w_2), ..., P(w_K)$
- 类条件概率分布函数的集合：$p(\mathbf{x}|w_m), k = 1, 2, ..., K$

每一类的后验概率可以计算如下：

$$P(w_j|\mathbf{x}) = \frac{p(\mathbf{x}|w_j)P(w_j)}{\sum_{k=1}^{K} p(\mathbf{x}|w_k)P(w_k)}, \quad j = 1, 2, ..., K$$

贝叶斯决策规则：

$$\hat{w} = \arg\max_{w_j} P(w_j|\mathbf{x}) = \arg\max_{w_j} p(w_j, \mathbf{x})$$



二分类规则将整个空间分为两个区域，将区域 $\mathcal{R}_1$ 分类为 w_1，区域 $\mathcal{R}_2$ 分类为 w_2。则出错的概率为：

$$
\begin{aligned}
P(error) &= P(\mathbf{x} \in \mathcal{R}_2, w_1) + P(\mathbf{x} \in \mathcal{R}_1, w_2) \\
&= P(\mathbf{x} \in \mathcal{R}_2 | w_1) P(w_1) + P(\mathbf{x} \in \mathcal{R}_1 | w_2) P(w_2) \\
&= \int_{\mathcal{R}_2} p(\mathbf{x}|w_1) P(w_1) d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|w_2) P(w_2) d\mathbf{x}
\end{aligned}
$$

一种艾滋病血液检测方法的检测精度如下表所示：

	患病	健康
阳性	99.9%	0.02%
阴性	0.1%	99.98%

已知甲所在的人群中约有 0.01% 携带艾滋病病毒。若甲进行此种血液检测检查得到的结果为阳性，应给出什么诊断结果（患病/健康）才能使诊断错误的概率最小？

$$
\begin{aligned}
P(\text{患病}|\text{阳性}) &= \frac{P(\text{阳性}|\text{患病})P(\text{患病})}{P(\text{阳性}|\text{患病})P(\text{患病}) + P(\text{阳性}|\text{健康})P(\text{健康})} \\
&= \frac{0.999 \times 0.0001}{0.999 \times 0.0001 + 0.0002 \times 0.9999} \\
&\approx 0.333 \\
P(\text{健康}|\text{阳性}) &\approx 0.667
\end{aligned}
$$

$$
P(\text{患病}|\text{阳性}) < P(\text{健康}|\text{阳性})
$$

Q: 最小化诊断错误概率是最合适的目标吗？

生成式模型与鉴别式模型

- 生成式：
  - 建模 $p(t,\mathbf{x}) = p(\mathbf{x}|t)P(t)$
  - 应用贝叶斯定理 $P(t|\mathbf{x}) = \dfrac{p(\mathbf{x}|t)P(t)}{p(\mathbf{x})}$
- 鉴别式：
  - 直接建模 $p(t|\mathbf{x})$

概率分布的监督学习中的参数估计

四个要素

- **数据**： 假定从基础分布生成
- **模型**： 结构和参数集， eg.
  - 高斯分布
  - 高斯混合模型
- **准则**
  - 最大似然
  - 最大参数后验
  - 鉴别性准则
- **假设**
  - 参数化分布表达式已知
  - 样本独立同分布（i.i.d.）

最大似然估计（MLE）

在 i.i.d. 假设下，模型参数 θ 关于样本集 $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2,...\mathbf{x}_n\}$ 的似然度为

$$p(\mathbf{X}|\theta)=\prod_{n=1}^{N}p(\mathbf{x}_n|\theta)$$

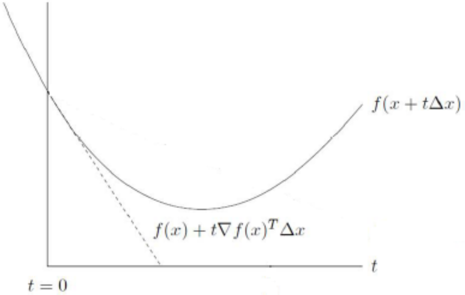
假设： θ 是确定的，尽管未知
目标： 寻找参数集 $\hat{\theta}$ 以最大化 $p(\mathbf{X}|\theta)$

实践中对数似然函数更易处理：

$$\mathcal{L}(\theta)=\log(p(\mathbf{X}|\theta))=\sum_{n=1}^{N}\log(p(\mathbf{x}_n|\theta))$$

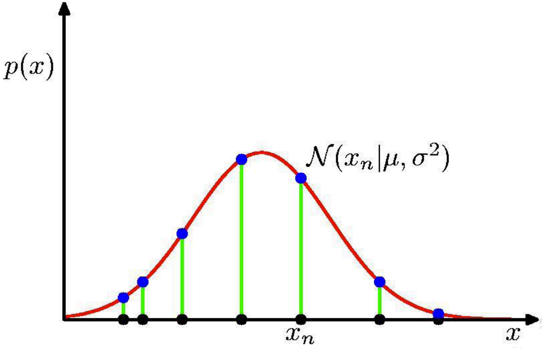
最大化似然度

一种典型的做法是寻找使梯度为 0 的 θ



$$\nabla_\theta \mathcal{L}(\theta) = \sum_{n=1}^{N} \nabla_\theta \log(p(\mathbf{x}_n|\theta)) = \sum_{n=1}^{N} \frac{\partial \log(p(\mathbf{x}_n|\theta))}{\partial \theta} = 0$$

高斯参数估计



$$p(\mathbf{x}|\theta) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu,\sigma^2)$$

考虑一维观测 $(d=1)$

$$\mathcal{L}(\theta)=\sum_{n=1}^{N}\log\left(p(x_n|\mu,\sigma)\right)=\sum_{n=1}^{N}\left(-\frac{1}{2}\log(2\pi\sigma^2)-\frac{(x_n-\mu)^2}{2\sigma^2}\right)$$

$$\begin{aligned}
\nabla_{\mu}\mathcal{L}(\theta) &= \frac{\partial\mathcal{L}(\mu)}{\partial\mu}=\sum_{n=1}^{N}(x_n-\mu)=0\\
&\Rightarrow \hat{\mu}=\frac{1}{N}\sum_{n=1}^{N}x_n\\
\nabla_{\sigma^2}\mathcal{L}(\theta) &= \frac{\partial\mathcal{L}(\sigma^2)}{\partial\sigma^2}=\sum_{n=1}^{N}\left(\frac{1}{\sigma^2}-\frac{(x_n-\hat{\mu})^2}{\sigma^4}\right)=0\\
&\Rightarrow \hat{\sigma}^2=\frac{1}{N}\sum_{n=1}^{N}(x_n-\hat{\mu})^2
\end{aligned}$$

向量微积分

向量微积分实用公式

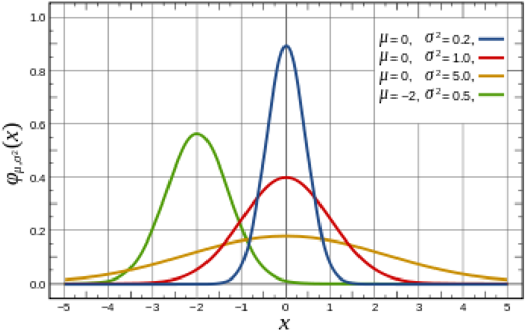
$$\frac{\partial}{\partial \boldsymbol{A}} \log |\boldsymbol{A}| = (\boldsymbol{A}^{-1})^\top$$
$$\frac{\partial}{\partial \boldsymbol{A}} (\mathbf{x}^\top \boldsymbol{A} \mathbf{y}) = \mathbf{x}\mathbf{y}^\top$$
$$\frac{\partial}{\partial \boldsymbol{x}} (\boldsymbol{x}^\top \boldsymbol{a}) = a$$
$$\frac{\partial}{\partial \boldsymbol{x}} (\boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x}) = (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x}$$

注:

$$|A| = \sum_{i=1}^{n} (-1)^{i+j} a_{ij} M_{ij} \qquad A^{-1} = \frac{1}{|A|} A^*$$

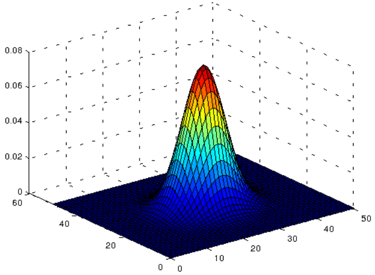
回顾：单高斯分布

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$



给定独立同分布数据 $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^\top$，则对数似然函数为

$$\log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$



充分统计量

$$\boldsymbol{\Gamma}_0 = \sum_{n=1}^{N} 1 = N$$

$$\boldsymbol{\Gamma}_1 = \sum_{n=1}^{N} \mathbf{x}_n$$

$$\boldsymbol{\Gamma}_2 = \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^\top$$

令对数似然函数的导数为 0，

$$\frac{\partial}{\partial \boldsymbol{\mu}} \log p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

解得

$$\hat{\boldsymbol{\mu}}_{\mathtt{ML}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n = \frac{\boldsymbol{\Gamma}_1}{\boldsymbol{\Gamma}_0}$$

同理有

$$\hat{\boldsymbol{\Sigma}}_{\mathtt{ML}} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathtt{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathtt{ML}})^\top = \frac{\boldsymbol{\Gamma}_2}{\boldsymbol{\Gamma}_0} - \hat{\boldsymbol{\mu}}_{\mathtt{ML}} \hat{\boldsymbol{\mu}}_{\mathtt{ML}}^\top$$

多项分布

1-of-K 编码: $\mathbf{x} = (0, 0, 1, 0, 0, 0)^\top$

$$P(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \forall k : \mu_k \geq 0 \text{ 且} \sum_{k=1}^{K} \mu_k = 1$$

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} P(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^\top = \boldsymbol{\mu} \qquad \sum_{\mathbf{x}} P(\mathbf{x}|\boldsymbol{\mu}) = \sum_{k=1}^{K} \mu_k = 1$$

最大似然参数估计:

$$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \quad P(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\left(\sum_n x_{nk}\right)} = \prod_{k=1}^{K} \mu_k^{m_k}$$

拉格朗日乘子

考虑优化问题

$$\begin{array}{ll} \text{最大化} & f(\mathbf{x}) \\ \text{满足} & g(\mathbf{x}) = 0 \end{array}$$

引入一个新变量（λ），称作**拉格朗日乘子**

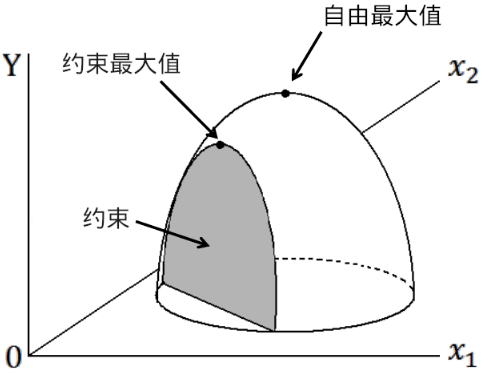
拉格朗日函数 $\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda \cdot g(\mathbf{x})$

解

$$\nabla_{\mathbf{x},\lambda} \mathcal{L}(\mathbf{x}, \lambda) = 0 \Leftrightarrow \begin{cases} \nabla_{\mathbf{x}} f(\mathbf{x}) = \lambda \nabla_{\mathbf{x}} g(\mathbf{x}) \\ g(\mathbf{x}) = 0 \end{cases}$$

约束优化可视化

图示为自由最优点与约束最优点之间的区别。



1-of-K 编码： $\mathbf{x} = (0,0,1,0,0,0)^\top$

$$P(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k} \qquad \sum_{k=1}^{K} \mu_k = 1$$

已知：

$$\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$$

$$P(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{\left(\sum_n x_{nk}\right)} = \prod_{k=1}^{K} \mu_k^{m_k}$$

使用拉格朗日乘子，

$$\sum_{k=1}^{K} m_k \log \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

$$\mu_k = -m_k/\lambda \qquad \mu_k^{\mathtt{ML}} = \frac{m_k}{N}$$

数据：

$$\mathbf{X} = \{x_1, \cdots, x_N\} \quad x_i > 0$$

模型：

$$p(x|\theta) = \begin{cases} \frac{1}{\theta_2 - \theta_1} & \theta_1 \leq x \leq \theta_2 \\ 0 & \texttt{otherwise} \end{cases}$$

准则：

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathbf{X}|\theta)$$

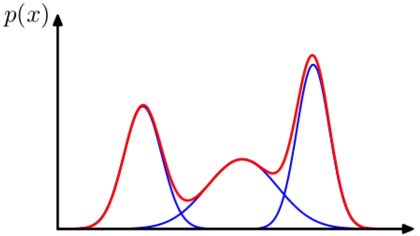
Q: θ 的最大似然估计是多少？

回顾：高斯分布的混合

将简单模型合并为一个复杂模型：

$$p(\mathbf{x}) = \sum_{k=1}^{K} \underbrace{\pi_k}_{\text{混合系数}} \underbrace{\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}_{\text{分量}}$$

$$\forall \pi_k > 0, \sum_k \pi_k = 1$$



$K = 3$

参数集：

- 混合权重：$\pi_1, \ldots, \pi_K$　　仅 $K-1$ 个自由参数
- 均值：$\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$
- 协方差：$\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$

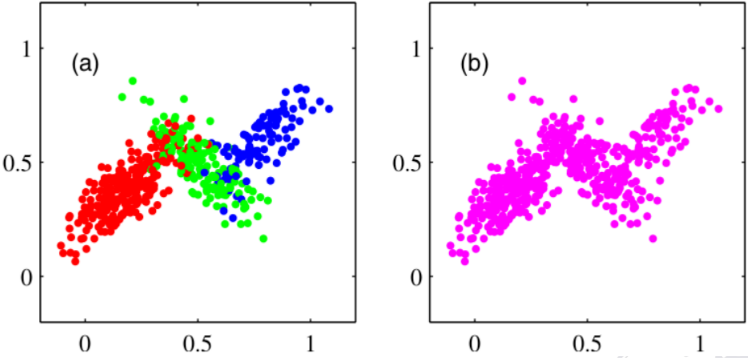
GMM 中的隐变量

$\mathbf{z} = (0, 0, \cdots, 1, \cdots, 0)$ 指示哪一高斯分量用于生成样本

称 $\mathbf{z}$ 为隐变量是因为它不能作为特征被观测到。根据 GMM 样本生成解释：

1. 选择一个高斯分量 $P(z_m = 1) = c_m$
2. 条件分布是高斯分布 $p(\mathbf{x}|z_m = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$
3. 因此，整个 GMM 为

$$p(\mathbf{x}) = \sum_{m=1}^{M} P(z_m = 1) p(\mathbf{x}|z_m = 1) = \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$$

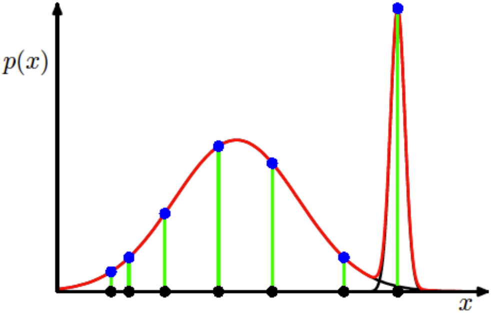


$\mathbf{z}$ 可被观测：记 z_n 表示样本 n 的已知高斯索引。

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \log\left(\sum_{m=1}^{M} \delta(z_n, m) c_m \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)\right) = \sum_{n=1}^{N} \log(c_{z_n} \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_{z_n}, \boldsymbol{\Sigma}_{z_n}))$$

GMM 参数估计的奇异性问题

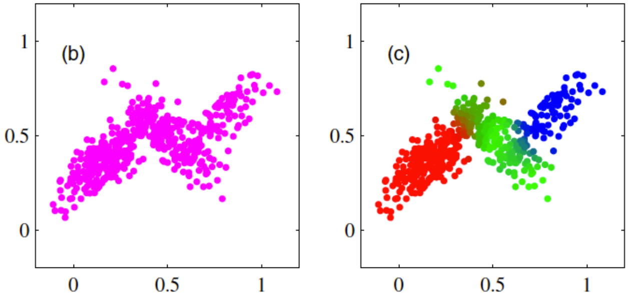
硬分配的问题



- z 无法被直接观测
- 假设 z 能被观测到会导致估计错误

GMM 中的隐变量

如果隐变量不能被直接观测



$\mathbf{z}$ **是隐变量**：由于 $\log \sum(\cdot)$ 的存在，直接优化会十分困难。

$$\mathcal{L}(\theta) = \sum_{n=1}^{N} \log \left(\sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}_n; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \right)$$

其中 $\theta = \{c_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m\}$ 是参数集合。

对隐变量进行软分配的直觉方法

- **困难**：我们不了解 x_n（以多大概率）属于哪一个高斯分量 m
- **解决方案**：
  1. 找个模型参数的初始值
  2. 对每一个样本 x_n 计算它属于每个高斯分量 m 的后验概率 $\gamma_m(n) = P(m|x_n)$，视其为观察到的“软性样本数量”
  3. 对于每个高斯分量 m，依据观察到的”软性样本数量“来做数据统计，得到每个高斯分量 m 的充分统计量

$$\boldsymbol{\Gamma}_0^m = \sum_{n=1}^{N} \gamma_m(n) \ , \quad \boldsymbol{\Gamma}_1^m = \sum_{n=1}^{N} \gamma_m(n)\mathbf{x}_n \ , \quad \boldsymbol{\Gamma}_2^m = \sum_{n=1}^{N} \gamma_m(n)\mathbf{x}_n\mathbf{x}_n^\top \ .$$

  4. 按照高斯参数重估公式得到每个高斯的参数

$$\boldsymbol{\mu}_m = \frac{\boldsymbol{\Gamma}_1^m}{\boldsymbol{\Gamma}_0^m} \ , \quad \boldsymbol{\Sigma}_m = \frac{\boldsymbol{\Gamma}_2^m}{\boldsymbol{\Gamma}_0^m} - \boldsymbol{\mu}_m\boldsymbol{\mu}_m^\top \ , \quad c_m = \frac{\boldsymbol{\Gamma}_0^m}{\sum_m \boldsymbol{\Gamma}_0^m}$$

  5. 返回第 2 步继续更新，直到似然度增加幅度足够小

间接最大似然估计

软分配直觉方法的数学支撑

- 从期望的视角重新整理准则函数：

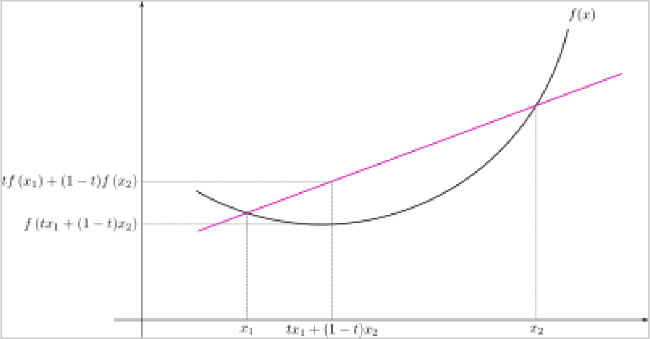
$$\mathcal{L}(\theta) = \sum_n \log p(\mathbf{x}_n|\theta) = \sum_n \log\left(\sum_m p(\mathbf{x}_n|m,\theta)P(m|\theta)\right)$$

- 间接估计需要使用一个**辅助函数**，它是准则函数的一个严格下界。

$$\begin{aligned}\mathcal{L}(\theta) &\geq \mathcal{Q}(\theta;\hat{\theta}) \\ \max_\theta \mathcal{L}(\theta) &\Leftrightarrow \max_\theta \mathcal{Q}(\theta;\hat{\theta})\end{aligned}$$

Jensen 不等式

若 X 是一个随机变量，φ 是一个凸函数，那么 $\varphi(\mathbb{E}[X]) \leq \mathbb{E}[\varphi(X)]$.



Jensen's inequality generalizes the statement that a secant line of a convex function lies above the graph.

凹函数则与之相反，因此有：

$$\log \mathbb{E}_{\mathbf{z}}[f(\mathbf{z})] \geq \mathbb{E}_{\mathbf{z}}[\log f(\mathbf{z})]$$

考虑已经存在一个初始参数 $\hat{\theta}$，因为 $\log$ 是凹函数，由 **Jensen 不等式**，有

$$
\begin{aligned}
\log p(\mathbf{X}|\theta) &= \sum_{n=1}^{N} \log \sum_{m=1}^{M} p(\mathbf{x}_n, m|\theta) \\
&= \sum_{n=1}^{N} \log \sum_{m=1}^{M} P(m|\mathbf{x}_n, \hat{\theta}) \frac{p(\mathbf{x}_n, m|\theta)}{P(m|\mathbf{x}_n, \hat{\theta})} \\
&\geq \sum_{n=1}^{N} \sum_{m=1}^{M} P(m|\mathbf{x}_n, \hat{\theta}) \log \frac{p(\mathbf{x}_n, m|\theta)}{P(m|\mathbf{x}_n, \hat{\theta})} \\
&= \sum_{n=1}^{N} \mathtt{H}\left(P(m|\mathbf{x}_n, \hat{\theta})\right) + \mathcal{Q}(\theta; \hat{\theta})
\end{aligned}
$$

将 $\log p(\mathbf{X}|\theta)$ 的下界定义为**辅助函数**：

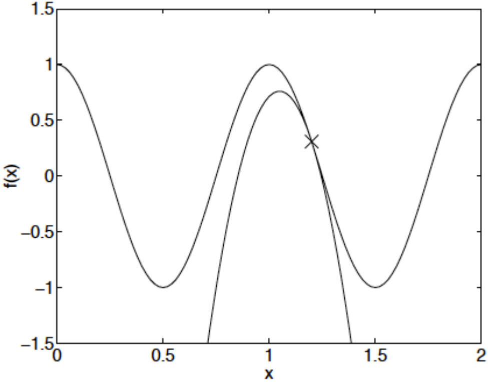
$$\mathcal{Q}(\theta;\hat{\theta}) = \sum_{n=1}^{N}\sum_{m=1}^{M} P(m|\mathbf{x}_n,\hat{\theta})\log\left(p(\mathbf{x}_n,m|\theta)\right)$$

在 GMM 的情况下，它的形式为

$$\begin{aligned}\mathcal{Q}(\theta;\hat{\theta}) &= K + \sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_m(n)\log c_m \\ &-\frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_m(n)\left(\log|\boldsymbol{\Sigma}_m| + (\boldsymbol{x}_n-\boldsymbol{\mu}_m)^\top\boldsymbol{\Sigma}_m^{-1}(\boldsymbol{x}_n-\boldsymbol{\mu}_m)\right)\end{aligned}$$

- $\gamma_m(n) = P(m|\mathbf{x}_n,\hat{\theta})$ 是 $\mathbf{x}_n$ 属于分量 m 的后验概率。
- 辅助函数 $\mathcal{Q}(\theta,\hat{\theta})$ 是完全数据集在 $\gamma_m(n)$ 上的期望。

似然度与辅助函数



$$
\begin{aligned}
\mathcal{L}(\theta) = \log p(\mathbf{X}|\theta) &\geq \mathcal{Q}(\theta;\hat{\theta}) \\
\mathcal{L}(\theta) - \mathcal{L}(\hat{\theta}) &\geq \mathcal{Q}(\theta;\hat{\theta}) - \mathcal{Q}(\hat{\theta};\hat{\theta})
\end{aligned}
$$

引理：Jensen 不等式在 $\hat{\theta}$ 处取等：

$$\mathcal{L}(\hat{\theta}) = \sum_{n=1}^{N} \mathtt{H}\left(P(m|\mathbf{x}_n,\hat{\theta})\right) + \mathcal{Q}(\hat{\theta};\hat{\theta})$$

则易得：

$$\begin{aligned}
\mathcal{Q}(\theta,\hat{\theta}) \quad &\geq \quad \mathcal{Q}(\hat{\theta},\hat{\theta}) \quad \Rightarrow \\
\sum_{n=1}^{N} \mathtt{H}\left(P(m|\mathbf{x}_n,\hat{\theta})\right) + \mathcal{Q}(\theta,\hat{\theta}) \quad &\geq \quad \sum_{n=1}^{N} \mathtt{H}\left(P(m|\mathbf{x}_n,\hat{\theta})\right) + \mathcal{Q}(\hat{\theta},\hat{\theta}) \quad \Rightarrow \\
\mathcal{L}(\theta) \quad &\geq \quad \mathcal{L}(\hat{\theta})
\end{aligned}$$

期望最大化（Expectation Maximization: EM）

- **期望（E 步骤）**：计算后验

$$\gamma_m(n) = P(m|\mathbf{x}_n, \hat{\theta}) = \frac{p(\mathbf{x}_n|m, \hat{\theta})P(m|\hat{\theta})}{\sum_{k=1}^{M} p(\mathbf{x}_n|k, \hat{\theta})P(k|\hat{\theta})}$$

- **最大化（M 步骤）**：寻找参数以最大化辅助函数

$$\begin{aligned}\mathcal{Q}(\theta; \hat{\theta}) = K &+ \sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_m(n) \log c_m \\ &- \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{M} \gamma_m(n) \left(\log|\boldsymbol{\Sigma}_m| + (\mathbf{x}_n - \boldsymbol{\mu}_m)^\top \Sigma_m^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_m)\right)\end{aligned}$$

GMM 的最大似然估计结果

给定如下充分统计量

$$\boldsymbol{\Gamma}_0^m = \sum_{n=1}^{N} \gamma_m(n) \ , \quad \boldsymbol{\Gamma}_1^m = \sum_{n=1}^{N} \gamma_m(n)\mathbf{x}_n \ , \quad \boldsymbol{\Gamma}_2^m = \sum_{n=1}^{N} \gamma_m(n)\mathbf{x}_n\mathbf{x}_n^\top \ .$$

GMM 的参数更新如下:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_m &= \frac{\boldsymbol{\Gamma}_1^m}{\boldsymbol{\Gamma}_0^m} \\
\hat{\boldsymbol{\Sigma}}_m &= \frac{\boldsymbol{\Gamma}_2^m}{\boldsymbol{\Gamma}_0^m} - \boldsymbol{\mu}_m\boldsymbol{\mu}_m^\top \\
\hat{c}_m &= \frac{\boldsymbol{\Gamma}_0^m}{\sum_m \boldsymbol{\Gamma}_0^m}
\end{aligned}$$

GMM 的 EM 算法过程

1. 设定初始参数 $\theta^{(0)}$，并令 $k=1$
2. E 步骤：使用 $\theta^{(k-1)}$ 为每个 $\mathbf{x}_n$ 计算后验

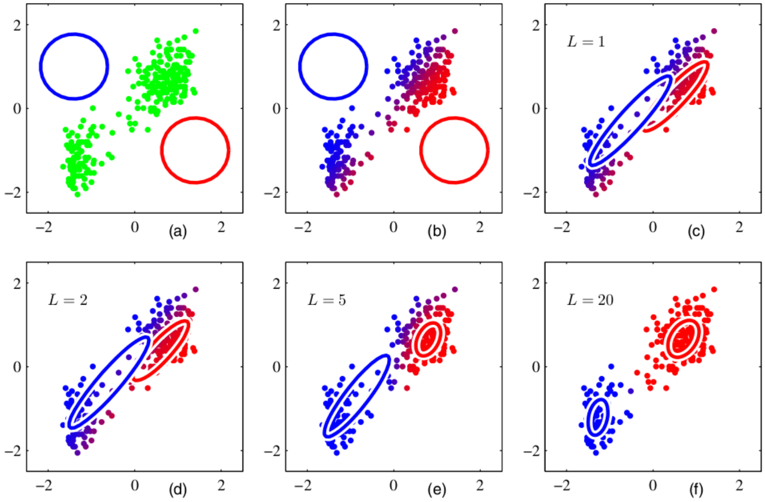
$$\gamma_m^{(k)}(n) = \frac{c_m^{(k-1)}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_m^{(k-1)}, \boldsymbol{\Sigma}_m^{(k-1)})}{\sum_{j=1}^{M} c_j^{(k-1)}\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j^{(k-1)}, \boldsymbol{\Sigma}_j^{(k-1)})}$$

3. M 步骤：使用上页的更新公式计算参数

$$c_m^{(k)}, \boldsymbol{\mu}_m^{(k)}, \boldsymbol{\Sigma}_m^{(k)}$$

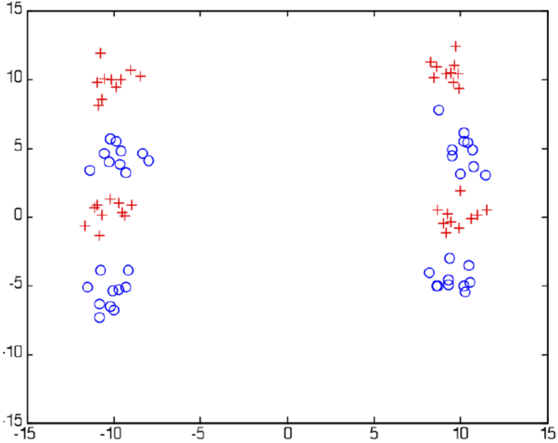
4. 令 $k=k+1$，并返回第 2 步，直至收敛

EM 示例

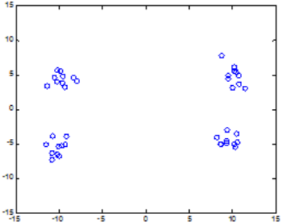


将 GMM 用于分类

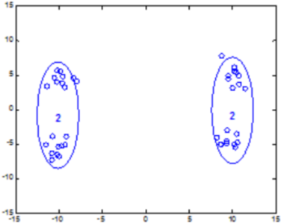
$$p(\mathbf{x}) = \sum_{k=1}^{2} \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$$



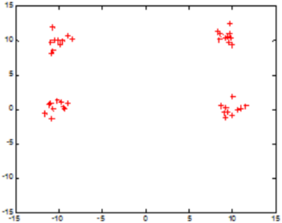
GMM 的最大似然训练



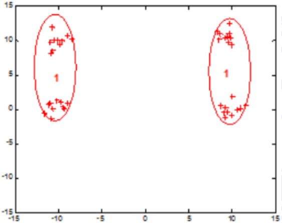
$$p(\mathbf{x}|l = b) = \sum_{k=1}^{2} \pi_k^b \mathcal{N}(\mathbf{x}|\mu_k^b, \Sigma_k^b)$$



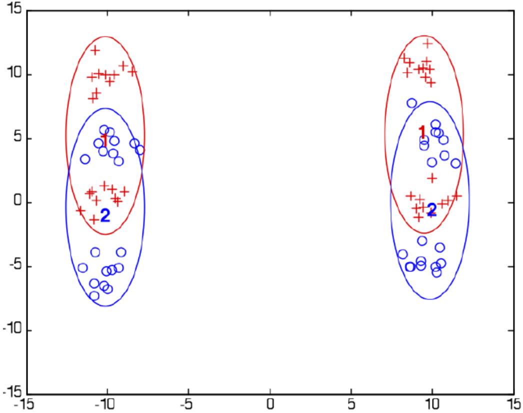
$$\mathcal{L}_{\texttt{ml}}(\theta_l) = \sum_n \log p(\mathbf{x}_n|l, \theta_l)$$



$$p(\mathbf{x}|l = r) = \sum_{k=1}^{2} \pi_k^r \mathcal{N}(\mathbf{x}|\mu_k^r, \Sigma_k^r)$$



使用 GMM 推理和决策



$$
\begin{aligned}
\hat{l} &= \arg\max_{l} P(l|\mathbf{x}) \\
&= \arg\max_{l} p(\mathbf{x}|l)P(l) \\
&= \arg\max_{l} P(l) \sum_{k=1}^{2} \pi_k^l \mathcal{N}(\mathbf{x}|\mu_k^l, \Sigma_k^l)
\end{aligned}
$$