



机器学习与数据挖掘

线性回归与分类的概率视角



课程大纲

- 介绍
- 从概率角度看回归
- 从概率角度看分类

条件概率视角

- 回归和分类的目标是在给定输入数据 x 的情况下，预测可能的输出 y

$$x \xrightarrow{\text{预测}} y$$

- 在回归和分类中，预测是通过确定性函数进行的

$$\text{Regression: } f(x) = \mathbf{x}w$$

$$\text{Classification: } f(x) = \sigma(\mathbf{x}w)$$

从概率的角度来看，为了在给定 x 的情况下预测输出 y ，我们只需要建模条件概率

$$p(y|x)$$

- 有了条件概率 $p(y|x)$, 输出可以按以下方式预测:

$$\text{均值: } \hat{y} = \int y p(y|x) dy$$

或

$$\text{最大后验: } \hat{y} = \arg \max_y p(y|x)$$

均值: 更侧重于预测输出的平均结果, 适用于需要连续预测或期望值估计的场景

最大后验: 更侧重于选择最有可能的类别或参数, 适用于分类问题和参数估计

回归和分类的目标都与条件概率紧密相关

课程大纲

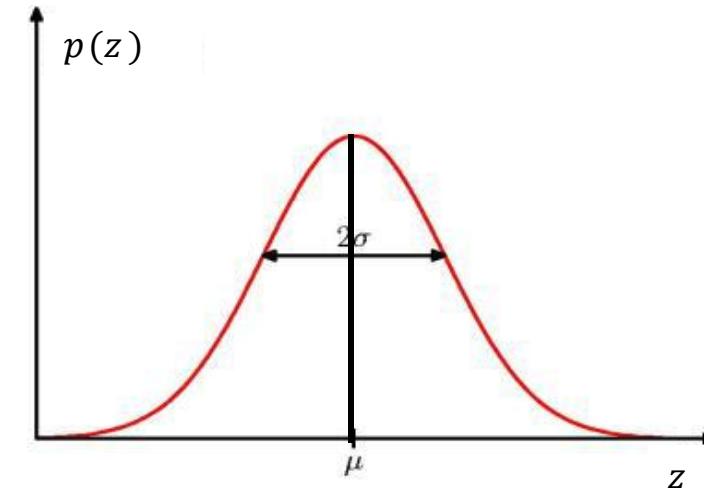
- 介绍
- 从概率角度看回归
- 从概率角度看分类

高斯分布

- 一元高斯分布

$$p(z) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(z - \mu)^2}{\sigma^2}\right] \triangleq \mathcal{N}(z; \mu, \sigma^2)$$

- μ 是均值
- $\sigma^2 = E[(z - \mu)^2]$ 是方差
- σ 是标准差



- μ 是分布的峰值和中心
- σ 决定了分布的离散程度（或宽度）

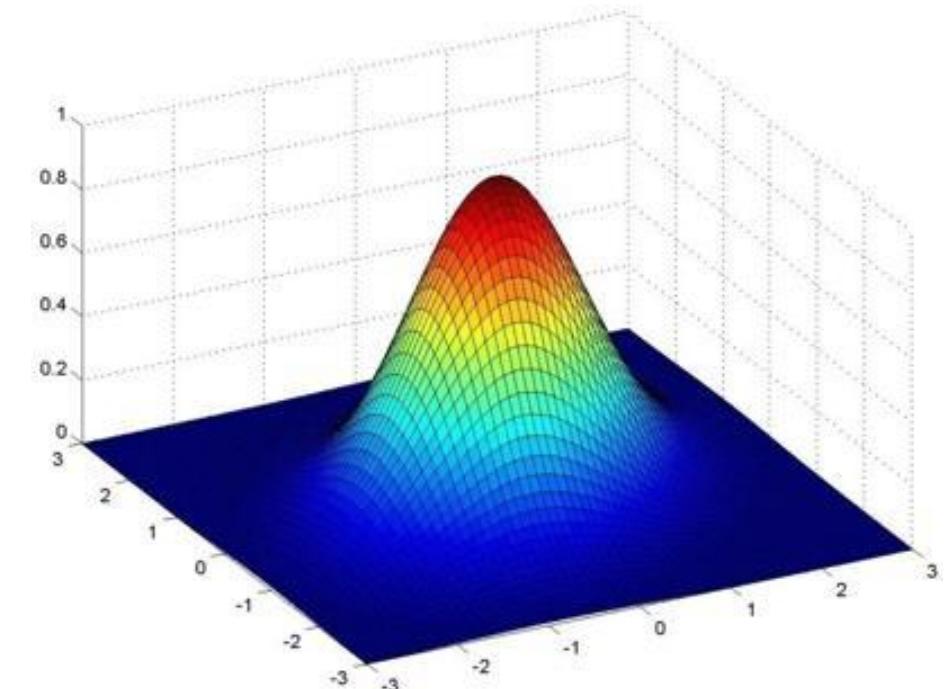
钟形

高斯分布

- 多元高斯分布

$$p(\mathbf{z}) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right\} \triangleq \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \Sigma)$$

- D 是维度
 - $\boldsymbol{\mu} \in R^D$ 是均值向量
 - $\Sigma \in R^{D \times D}$ 是标准差
- $\boldsymbol{\mu}$ 是分布的**峰值**和**中心**
- Σ 决定了分布的离散程度 (或宽度)

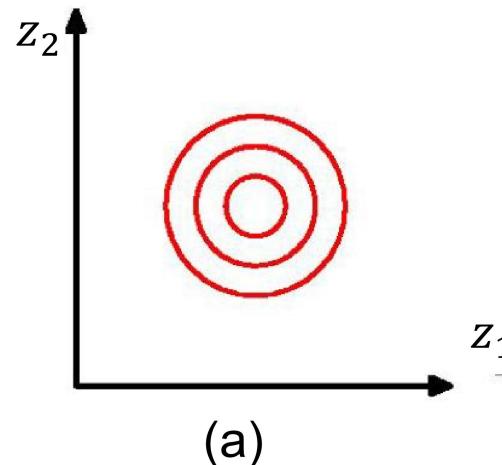


高斯分布

- 不同 Σ 类型下的分布形状

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

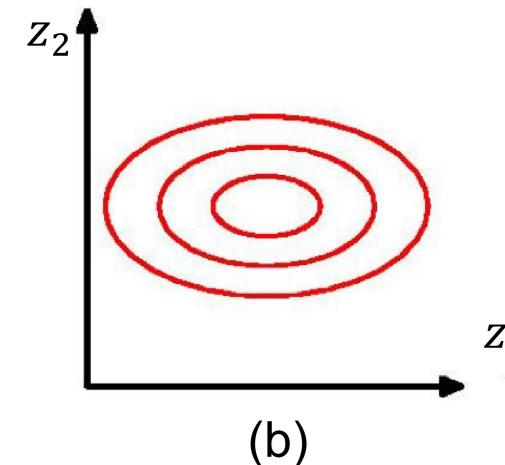
$$\sigma_1^2 = \sigma_2^2$$



(a)

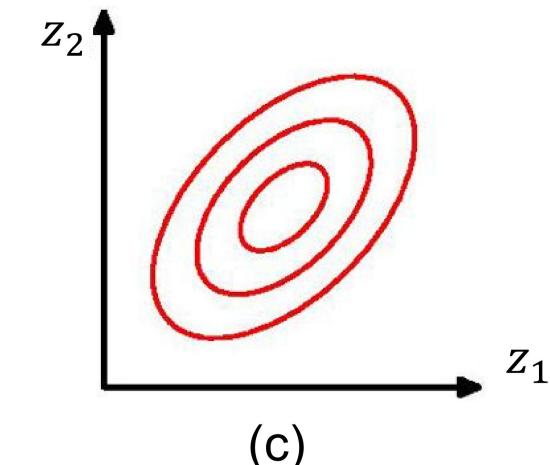
$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$\sigma_1^2 > \sigma_2^2$$



(b)

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho \\ \rho & \sigma_2^2 \end{bmatrix}$$



(c)

- 无论 Σ 如何变化, 峰值总是位于 μ (单峰)

高斯分布

- 对于每一个协方差矩阵 Σ ，可以分解为 $\Sigma = \mathbf{U}\Lambda\mathbf{U}^T$
 - \mathbf{U} 是一个正交矩阵，其中 $\mathbf{U}\mathbf{U}^T = \mathbf{I}$
 - Λ 是一个对角矩阵
- 使 $\mathbf{z}' = \mathbf{U}^T \mathbf{z}$ ， $\boldsymbol{\mu}' = \mathbf{U}^T \boldsymbol{\mu}$ ，分布可以表示为：

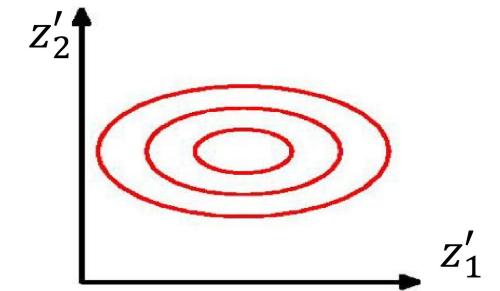
$$p(\mathbf{z}') = \frac{1}{(2\pi)^{D/2} |\Lambda|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{z}' - \boldsymbol{\mu}')^T \Lambda^{-1} (\mathbf{z}' - \boldsymbol{\mu}') \right\}$$

高斯分布

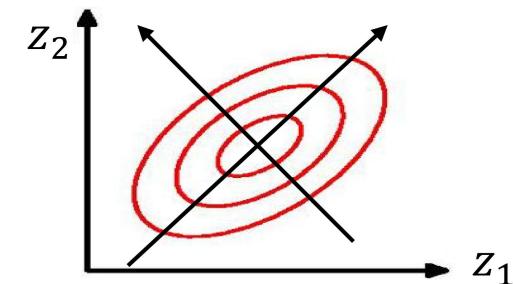
- 使 $\mathbf{z}' = \mathbf{U}^T \mathbf{z}$, $\boldsymbol{\mu}' = \mathbf{U}^T \boldsymbol{\mu}$, 分布可以表示为:

$$p(\mathbf{z}') = \frac{1}{(2\pi)^{D/2} |\Lambda|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{z}' - \boldsymbol{\mu}')^T \Lambda^{-1} (\mathbf{z}' - \boldsymbol{\mu}') \right\}$$

因此 $p(\mathbf{z}')$ 的形状看起来是:



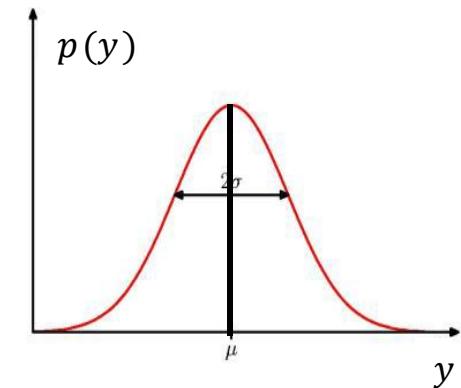
而 $p(\mathbf{z})$ 的形状被 \mathbf{U} 旋转了:



线性回归

- 从概率视角来看，为了进行预测，我们只需要指定条件概率分布 $p(y|x)$ 。对于回归问题，我们假设这个分布是正态分布

$$\begin{aligned} p(y|x; \mathbf{w}) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(y - \mathbf{x}\mathbf{w})^2}{\sigma^2}\right] \\ &= \mathcal{N}(y; \mathbf{x}\mathbf{w}, \sigma^2) \end{aligned}$$



- 我们使用分布的均值进行预测，即：

$$\hat{y} = \mathbf{x}\mathbf{w}$$

在这里得到的 \mathbf{w} 与传统回归中得到的 \mathbf{w} 相同吗？

- 模型训练的目标是找到使对数概率最大化的参数 \mathbf{w} , 即:

$$\max_{\mathbf{w}} \log p(y|\mathbf{x}; \mathbf{w})$$

对数似然函数

- 从 $p(y|\mathbf{x}; \mathbf{w})$ 的表达式中, 我们得到:

$$\log p(y|\mathbf{x}; \mathbf{w}) = -\frac{1}{2} \frac{(y - \mathbf{x}\mathbf{w})^2}{\sigma^2} + \text{constant}$$

因此, 最大化对数似然 $p(y|\mathbf{x}; \mathbf{w})$ 等价于最小化

$$\min_{\mathbf{w}} (y - \mathbf{x}\mathbf{w})^2,$$

这与回归中使用的损失函数是相同的

- 对于 N 个训练样本 $(\mathbf{x}^{(i)}, y^{(i)})$, 通过假设它们是 **独立同分布 (i.i.d.)** 的, 可以得到它们的联合条件概率密度函数:

$$p(y^{(1)}, \dots, y^{(N)} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2} \frac{(y^{(i)} - \mathbf{x}^{(i)} \mathbf{w})^2}{\sigma^2}\right]$$

- 对数似然函数是:

$$\log p(y^{(1)}, \dots, y^{(N)} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}) = -\frac{1}{2\sigma^2} \sum_{i=1}^N (y^{(i)} - \mathbf{x}^{(i)} \mathbf{w})^2 + \text{constant}$$

- 最大化对数似然 $\log p(y^{(1)}, \dots, y^{(N)} | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$ 等价于最小化

$$L(\mathbf{w}) = \sum_{i=1}^N (y^{(i)} - \mathbf{x}^{(i)} \mathbf{w})^2,$$

你能想出为什么这里会出现求和符号 Σ 吗?

这与回归中使用的损失函数是相同的

- 从概率建模的角度来看，线性回归实际上等价于：

- 建模：假设条件分布是对角高斯分布
 - 训练：通过最大化对数似然来训练模型

课程大纲

- 介绍
- 从概率角度看回归
- 从概率角度看分类

伯努利分布

- 伯努利分布与其他分布的关系

- 与二项分布的关系:

伯努利分布是二项分布的一种特殊情况

- 与高斯分布的关系:

在某些情况下，伯努利分布可以视为高斯分布的一个特例，
特别是当高斯分布的协方差矩阵是对角矩阵时

这是因为在这种情况下，高斯分布退化为伯努利分布，其中协方差矩阵的对角线元素表示变量间不相关，即每个变量独立地取值为 0 或 1

伯努利分布

- 伯努利分布

$$p(z) = \begin{cases} \pi, & \text{if } z = 1 \\ 1 - \pi, & \text{if } z = 0 \end{cases}$$

其中 $\pi \in [0, 1]$ 是 z 等于 1 的概率

- $p(z)$ 可以简洁地表示为:

$$p(z) = \pi^z \cdot (1 - \pi)^{1-z}$$

其中 $z = 0$ 或 1

二分类

- 为了实现二元分类，我们假设条件概率是伯努利分布

$$p(y|x) = (\sigma(xw))^y \cdot (1 - \sigma(xw))^{1-y}$$

其中 $\pi = \sigma(xw)$; 且 $y = 0$ 或 1

- 训练目标是最大化对数似然函数

$$\log p(y|x) = y \log \sigma(xw) + (1 - y) \log (1 - \sigma(xw))$$

二分类

回想一下，逻辑回归最小化：

cross entropy

$$\triangleq -y \log \sigma(\mathbf{x} \mathbf{w}) - (1 - y) \log (1 - \sigma(\mathbf{x} \mathbf{w}))$$

最大化 $\log p(y|\mathbf{x})$ 等价于最小化交叉熵

- 逻辑回归等价于：
 - 建模：假设输出服从伯努利条件分布
 - 训练：通过最大化对数似然来训练模型

类别分布

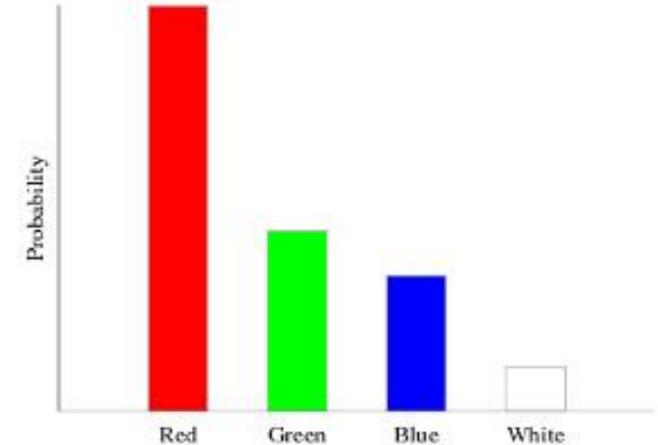
- 类别分布

$$p(\mathbf{z} = \text{one hot}_k) = \pi_k$$

- 其中 $\text{one hot}_i = [0, \dots, 0, 1, 0, \dots, 0]$ 是一个独热向量，其中第 i 个元素是唯一非零的元素 1
- $\sum_{k=1}^K \pi_k = 1$
- 该分布可以等价地写为：

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$$

其中 \mathbf{z} 是一个独热向量



多分类

- 建模：通过将概率 π_k 设置

$$\pi_k = \text{softmax}_k(\mathbf{x} \mathbf{W}),$$

假设条件概率分布服从类别分布

$$p(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^K [\text{softmax}_k(\mathbf{x} \mathbf{W})]^{y_k}$$

多分类

条件概率分布服从类别分布: $p(\mathbf{y}|\mathbf{x}) = \prod_{k=1}^K [\text{soft max}_k(\mathbf{x}\mathbf{W})]^{y_k}$

- 训练: 给定一个训练样本 (\mathbf{x}, \mathbf{y}) , 模型通过最大化对数似然函数进行训练

$$\begin{aligned} & \log p(\mathbf{y}|\mathbf{x}) \\ &= \sum_{k=1}^K y_k \cdot \log (\text{soft max}_k(\mathbf{x}\mathbf{W})) \\ & \quad \textcolor{red}{=- \text{cross entropy}} \end{aligned}$$

总结

- 回归、逻辑回归和多类别回归可以被归纳到同一个通用框架下

1) 建模: 为输出 y 假设不同的条件概率密度函数

- 回归: 高斯分布
- 逻辑回归: 伯努利分布
- 多类别逻辑回归: 分类分布

2) 训练: 最大化对数似然函数