

模式识别与机器学习

期末复习

➤ 模式识别与机器学习

- 模式识别 vs. 机器学习：一体两面，不做区分
- 模式识别：利用计算机对物理对象进行分类，在风险最低的条件下，使识别的结果尽量与真实情况相符合
 - 在特征空间和解释空间之间找到一种映射关系： $y = f(x)$
- 机器学习：利用大量的训练数据，获得产生数据的模式 f 并进行预测

➤ 风险和错误率

■ 错误率: $P(error) = \int P(error|\mathbf{x})p(\mathbf{x})d\mathbf{x}$

- 无需对 \mathbf{x} 积分, 对每个 \mathbf{x} 都求错误率最小的决策, 就是求对所有 \mathbf{x} 平均错误率最小的决策
- $P(error|\mathbf{x}) = 1 - P(Y = y|\mathbf{x})$

■ 风险: 平均损失

■ 损失 L_{cy} : 将本应属于 y 类的模式判别成属于 c 类的代价

$$R(\hat{y}(\mathbf{x}) = c|\mathbf{x}) = \sum_{y=1}^C L_{cy}P(Y = y|\mathbf{x})$$

➤ 机器学习的3个方面

- 函数族 \mathcal{F} : $f \in \mathcal{F}$
- 目标函数 $J(f)$: 度量 f 的好坏
- 优化算法: $f^* = \arg \min_f J(f)$

➤ 1. 函数族

- 线性模型: $f(x) = w^T x$
- 多项式
- 核方法 (由核函数决定)
- 决策树 (分段常数)
- 神经网络 (由网络结构决定)
- 集成学习: 多棵决策树的加权平均
 - 随机森林
 - GBDT
- 概率图模型: 利用条件独立假设, 简化概率计算
- 参数模型: 线性模型、多项式、神经网络、概率图模型
- 非参数模型: 核方法、决策树、集成学习

➤ 2. 目标函数

■ $J(f) = L(f) + \lambda R(f)$

■ 损失函数 $L(f)$: 函数 f 与训练数据的拟合程度

- 负log似然损失: L2/L1损失、交叉熵损失
- 合页损失 (SVM)、 ϵ 不敏感损失 (SVR)
- 聚类损失: K均值、谱聚类
- 降维损失: 结构保持、重构

■ 正则项 $R(f)$: 函数 f 自身的复杂程度

- L2正则、L1正则
- K-Lipschitz 连续 (WGAN)

■ 正则参数 λ

- λ 越小, 模型越复杂, 训练误差越小, 偏差小、方差大
- λ 越大, 模型越简单, 训练误差越大, 偏差大、方差小

➤ 模型的其他复杂度参数

- 线性模型：特征的数目
- 多项式：多项式阶数
- 核方法：核函数超参数（RBF核的核函数宽度）
- 决策树（树的最大深度、叶子结点数目、叶子结点代表的训练样本数目、...）
- 神经网络
 - 网络的连接方式：全连接、局部连接、...
 - 网络的层数
 - 每层的神经元的数目

➤ 3. 优化算法

■ $f^* = \arg \min_f J(f)$

■ 梯度下降/上升

- 梯度:

 - 常用损失函数的梯度计算

 - 批处理梯度下降、随机梯度下降、小批量梯度下降

 - 动量法

- 学习率

 - 学习率的影响

 - 自适应学习率

■ 坐标下降/上升: 每次选择一个或一部分参数更新

- K均值聚类

- EM

- Lasso、SMO

➤ 第2章 生成式分类器

■ 生产式分类器: $P(y = c|\mathbf{x}) = \frac{p(\mathbf{x}|y = c)P(y = c)}{\sum_{c'} p(\mathbf{x}|y = c')P(y = c')}$

■ 类先验 $P(y = c)$

- 两类: $y \sim \text{Bernoulli}(\theta)$
- 多类: $y \sim \text{Multinoulli}(\boldsymbol{\theta})$

■ 类条件 $p(\mathbf{x}|y = c)$

- 高斯判别分析: $p(\mathbf{x}|y = c) = N(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$

- 朴素贝叶斯: $p(\mathbf{x}|y = c) = \prod_{j=1}^D p(x_j|y = c)$

贝叶斯分类器判别函数

- 深度生成模型: 学习从简单分布 (高斯) 到 $p(\mathbf{x}|y = c)$ 的映射 (神经网络)

➤ 高斯判别分析

$$\blacksquare P(Y = c|\mathbf{x}) = \frac{p(\mathbf{x}|Y = c)P(Y = c)}{\sum_{c'} p(\mathbf{x}|Y = c')P(Y = c')}$$

$$p(\mathbf{x}|Y = c) = \frac{1}{(2\pi)^{\frac{D}{2}} |\boldsymbol{\Sigma}_c|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1}(\mathbf{x} - \boldsymbol{\mu}_c)\right)$$

$$\blacksquare \text{类别 } c \text{ 的判别函数: } f_c(\mathbf{x}) = -\frac{1}{2} \ln(|\boldsymbol{\Sigma}_c|) - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1}(\mathbf{x} - \boldsymbol{\mu}_c) + \ln P(Y = c)$$

■ 当 $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$ 时,

$$\begin{aligned} f_1(\mathbf{x}) - f_2(\mathbf{x}) &= -\frac{1}{2} \mathbf{x}^T (\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}) \mathbf{x} + (\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} - \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1}) \mathbf{x} + b + \ln \frac{P(Y = 1)}{P(Y = 2)} \\ &= (\boldsymbol{\mu}_1^T - \boldsymbol{\mu}_2^T) \boldsymbol{\Sigma}^{-1} \mathbf{x} + b + \ln \frac{P(Y = 1)}{P(Y = 2)} \end{aligned} \quad b = -\frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2$$

第3章 判别式分类器

■ 函数族：线性函数 $f(x) = w^T x$

■ 目标函数&优化算法

名称	目标函数	优化算法
Fisher判别分析	$J(w) = \frac{w^T S_b w}{w^T S_w w}$	解析解 $w = S_w^{-1}(\mu_1 - \mu_2)$
感知器	$J(w) = \sum_{x_i \in \mathcal{M}} -y_i f(x_i)$	随机梯度下降

■ 两类Fisher判别的判别面为： $w^T x = \frac{1}{2} w^T (\mu_1 + \mu_2)$

• 投影后的类中心连线的垂直平分线为判别面 $(\mu_1 - \mu_2)^T S_w^{-1} x = \frac{1}{2} (\mu_1 - \mu_2)^T S_w^{-1} (\mu_1 + \mu_2)$

➤ 第3章 判别式分类器

■ 决策树：分段常数函数

■ 损失函数：

- 分类：Gini指数
- 回归：L2损失

■ 正则：

- 叶子结点数目

■ 优化算法：贪心

- 精确穷举搜索
- 直方图近似搜索

➤ 第5章 线性回归

■ 函数族：线性函数 $\hat{y} = f(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$

■ 损失函数

• L2损失： $L(\hat{y}, y) = \frac{1}{2} (\hat{y} - y)^2$

• Huber损失： $L_\delta(\hat{y}, y) = \begin{cases} \frac{1}{2} (\hat{y} - y)^2 & |r| \leq \delta \\ \delta |\hat{y} - y| - \frac{1}{2} \delta^2 & otherwise \end{cases}$

■ 优化算法

- 解析法
- 梯度下降
- 坐标下降 (了解)

$$\frac{\partial L}{\partial \mathbf{w}} = \frac{\partial L}{\partial f} \frac{\partial f}{\partial \mathbf{w}} = (\hat{y} - y) \mathbf{x}$$

第6章 Logistic回归

■函数族：线性函数

$$\mu(\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x})$$

■损失函数

$$p(y|\mathbf{x}; \mu) = \mu(\mathbf{x})^y (1 - \mu(\mathbf{x}))^{(1-y)}$$

- 交叉熵损失/负log似然损失

$$L(\mu(\mathbf{x}), y) = -y \ln(\mu(\mathbf{x})) - (1 - y) \ln(1 - \mu(\mathbf{x}))$$

■优化算法

- 梯度下降
- 牛顿法 (了解)

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= \frac{\partial L}{\partial \mu} \frac{\partial \mu}{\partial \mathbf{w}} = \left(-\frac{y}{\mu(\mathbf{x})} + \frac{1-y}{1-\mu(\mathbf{x})} \right) \mu(\mathbf{x})(1-\mu(\mathbf{x})) \frac{\partial(\mathbf{w}^T \mathbf{x})}{\partial \mathbf{w}} \\ &= (\hat{y} - y) \mathbf{x} \end{aligned}$$

■多类分类Logistic回归 (softmax分类器)：了解

➤ 第7章 SVM

■ 样本点的（函数）间隔： $y(\mathbf{w}^T \mathbf{x} + b)$

■ 分类器的间隔： $\frac{2}{\|\mathbf{w}\|_2}$

■ 线性SVM

- 硬间隔/软间隔
- 支持向量
- 合页损失
- 原问题 vs. **对偶问题**

■ 核化SVM

- 常用核函数及其复杂度参数

■ SVR

第7章 SVM

■ 硬线性SVM

• 原问题

$$\min \frac{1}{2} \|\mathbf{w}\|_2^2$$

$$\text{s. t. } y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \dots, N$$

• 对偶问题

$$\max \left(\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right)$$

$$\text{s. t. } \sum_{i=1}^N \alpha_i y_i = 0$$

$$0 \leq \alpha_i, \quad i = 1, 2, \dots, N$$

第7章 SVM

■ 硬线性SVM

• 对偶问题

$$\begin{aligned} \max & \left(\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right) \\ \text{s. t. } & \sum_{i=1}^N \alpha_i y_i = 0 \\ & 0 \leq \alpha_i, \quad i = 1, 2, \dots, N \end{aligned}$$

• 支持向量: $\alpha_i \neq 0$

• 对偶解 \rightarrow 原问题解

$$\mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i$$

用任意一个支持向量即可求得 b : $b = y_i - \mathbf{w}^T \mathbf{x}_i$

第7章 SVM

■ 软间隔SVM

合页损失: $\xi = L_{Hinge}(y, \hat{y}) = \begin{cases} 0 & y\hat{y} \geq 1 \\ 1 - y\hat{y} & \text{otherwise} \end{cases}$

• 原问题 $J(\mathbf{w}, b, C) = C \sum_{i=1}^N \xi_i + \frac{1}{2} \|\mathbf{w}\|_2^2$

$$\text{s. t. } y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, N$$

$$\xi_i \geq 0, \quad i = 1, 2, \dots, N$$

• 对偶问题 $\max \left(\sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \right)$

$$\text{s. t. } \sum_{i=1}^N \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, N$$

➤ 第4章 特征选择和提取

■ 数据预处理

- 常用数值型特征的预处理/编码方案：log变换、量化
- 常用类别型特征的编码方案：**独热编码**、计数编码、hash编码、嵌入编码
- 特征缩放：**标准化**、正规化

■ 特征提取：重点掌握 **PCA**

■ 特征选择：了解

■ 模型训练：给定训练样本 $X = (x_1, x_2, \dots, x_N)$

- 1. 样本中心化: $\bar{x} = \frac{1}{N} \sum_{j=1}^N x_j$, $x_i = x_i - \bar{x}$
- 2. 计算 $S = XX^T$
- 3. 对 S 做特征值分解
- 4. D' 最大特征值对应的特征向量: $w_1, w_2, \dots, w_{D'}$ (D' 的选择)

输出: $W = (w_1, w_2, \dots, w_{D'})$

■ 推理/测试：对新的样本 x ，其对应的低维表示为: $z = W^T(x - \bar{x})$

■ 重构：给定 z ，重构结果为: $\hat{x} = Wz + \bar{x}$

➤ 第8章 统计学习基础

■ 经验风险

■ 期望风险

- 结构风险：经验风险 + 正则

■ 训练误差、验证误差、测试误差

- 交叉验证

■ 过拟合、欠拟合

■ 泛化误差分解：偏差-方差分解

■ 泛化误差上界

- 训练误差 $E_{\text{train}}(f^*)$ 、训练样本数目 N 、VC维 d_{vc}

$$E_{\text{train}}(f^*) - \Omega(d_{\text{vc}}, N, \delta) \leq E_{\text{test}}(f^*) \leq E_{\text{train}}(f^*) + \Omega(d_{\text{vc}}, N, \delta)$$

$$\Omega(d_{\text{vc}}, N, \delta) = \sqrt{\frac{8}{N} \ln \left(4 \frac{(2N)^{d_{\text{vc}}} + 1}{\delta} \right)}$$

➤ 训练集、验证集、测试集

■ 验证集：选择模型

验证集	训练集	测试集
-----	-----	-----

- 用样本外误差，估计测试误差
- 验证误差是真实误差的无偏估计，两者的差距与验证集的大小成反比
- 用不同超参数在验证集上的性能做模型选择（确定最佳超参数）
- 确定超参数后，用全体训练数据（训练集+验证集）训练模型参数
- 当训练数据集较小时，可采用交叉验证方式得到验证集

■ 测试集：评估模型

- 测试集是从总体选出来的部分样本，与训练集不重合，模拟没有见过但未来可能遇到的数据
- 用测试误差估计真实误差

➤ 第9章 集成学习

■ Bagging: 降低方差、偏差不变

- 基学习器类型相同: 通常较**复杂**, 偏差较小, 方差较大
- 通过对样本和特征进行随机采样得到不同的训练集训练各基学习器
- 模型融合: 多个基学习器结果的平均/投票
- 多个基学习器可并行训练

■ Boosting: 降低偏差、方差不变

- 基学习器类型相同: 通常较**简单**, 偏差较大, 方差较小
- 每次迭代通过改变样本的权重 (AdaBoost) 或改变标签 (GBM) 得到不同的训练集训练各基学习器、样本和特征可随机采样
- 模型融合: 多个基学习器结果的加权平均
- 多个基学习器顺序训练, 不能并行

■ Stacking: 了解

- 基学习器类型可以不同

➤ 第10章 聚类

■ 聚类性能评价指标：了解

- 簇内距离越小越好，簇间距离越大越好

■ 聚类算法

● K均值聚类

- 高斯混合模型
 - EM算法：了解
- 层次聚类：了解
- 基于密度的聚类：DBSCAN
- 基于图的聚类：了解

➤ 第11章 降维

■ 线性降维

- PCA
- 多维尺度缩放 (MDS) : 了解

■ 非线性降维: 至少掌握1种算法

- 重构残差最小: 核化PCA、**自编码器**
- 全局距离保持: ISOMAP 了解
- 邻域距离保持: Laplacian Eigenmaps、T-NSE, UMAP了解

➤ 第12章 半监督学习

■ 半监督学习的三个基本假设

- 高密度区域平滑假设
- 聚类假设/低密度分隔假设
- 流形假设

■ 半监督学习算法：至少熟悉一种算法原理

- 自我训练
- 多视角学习
- 生成模型
- 半监督SVM
- 基于图的算法
- 半监督聚类

➤ 第13章 人工神经网络

■ 神经元的结构

- **激活函数：非线性**、梯度消失、计算简单
- ReLU：最常用
- sigmoid：主要用于门控函数
- tanh：主要用于RNN

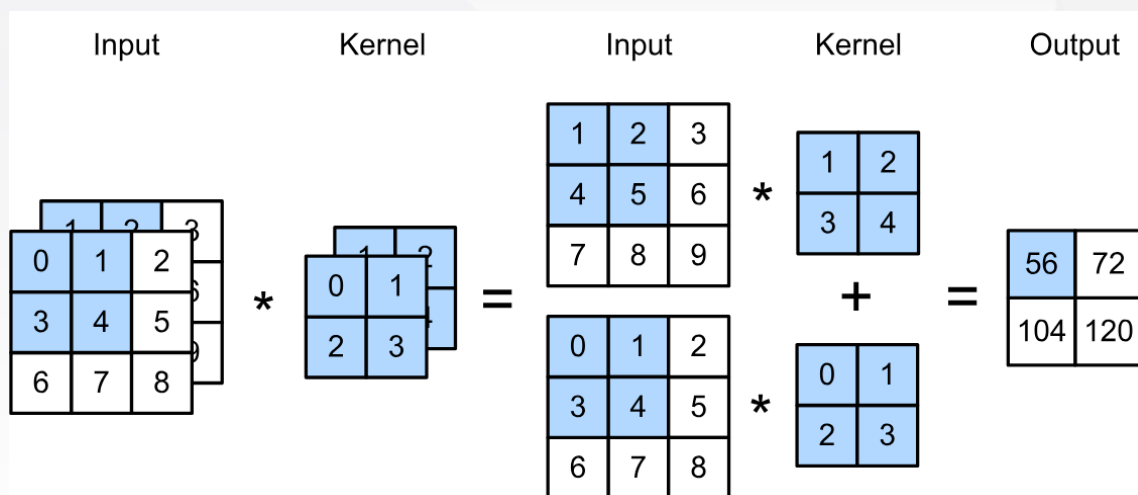
■ 神经网络结构

- 全连接
- **卷积**：局部连接、权值共享
- 循环：历史信息压缩、梯度消失、梯度爆炸
- **跳跃连接**：缓减梯度消失
- Transformer

>> 卷积

■根据数据特点设计的网络结构：局部连接、权值共享

- 平移不变、模式局部相关



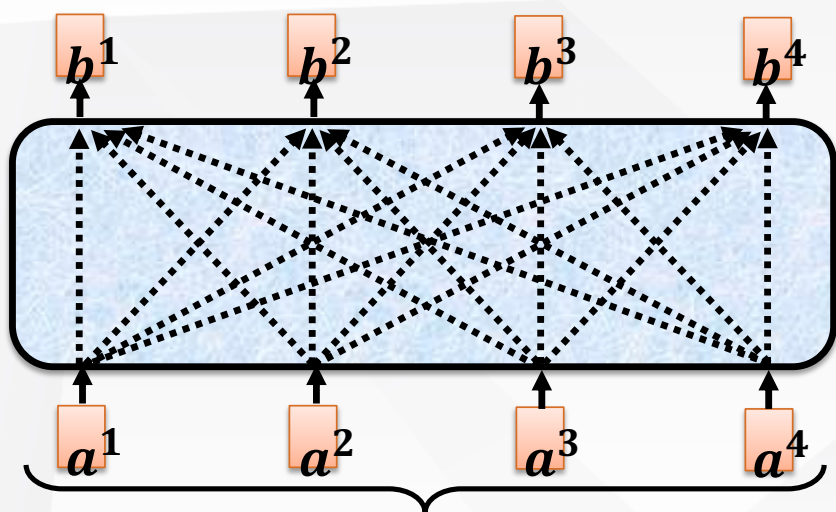
$$(1 \times 1 + 2 \times 2 + 4 \times 3 + 5 \times 4) + (0 \times 0 + 1 \times 1 + 3 \times 2 + 4 \times 3) = 56$$

- 输入： $X: c_i \times n_h \times n_w$ ，输出： $Y: c_o \times m_h \times m_w$ ，卷积核大小为 $k_h \times k_w$ ，则卷积核的参数数目为（不考虑偏置项 b ）：

$$c_o \times c_i \times k_h \times k_w$$

Transformer

■ 自注意力

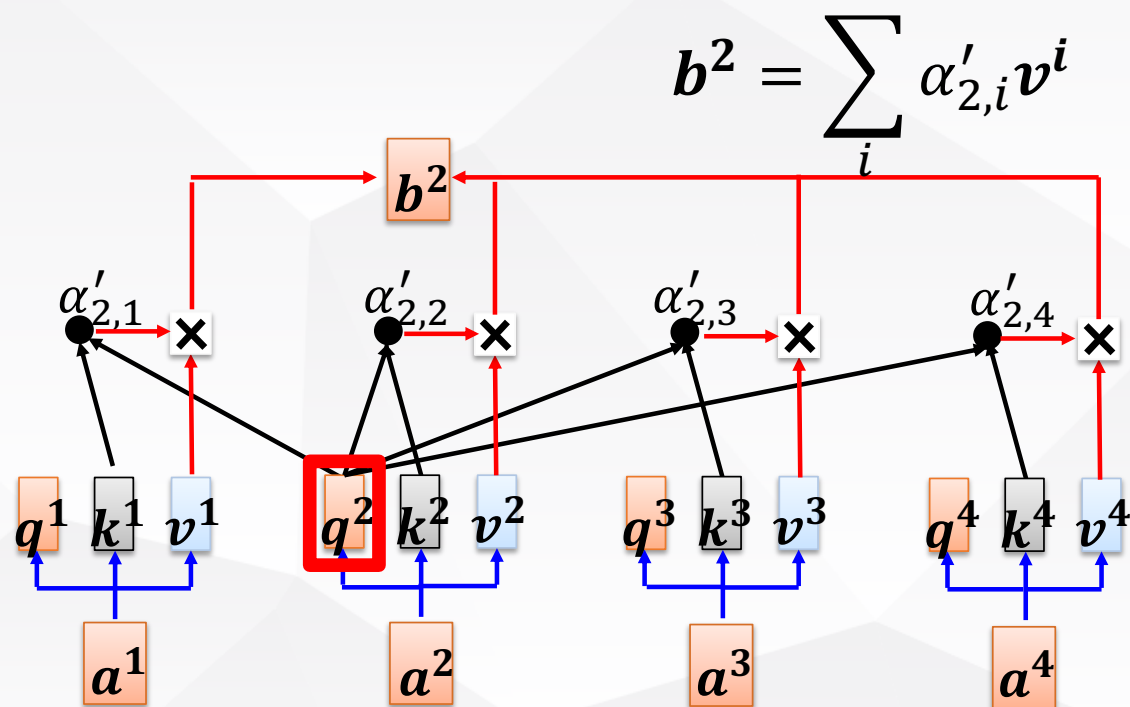


输入或隐含层

$$Q = IW^q$$

$$K = IW^k$$

$$V = IW^v$$

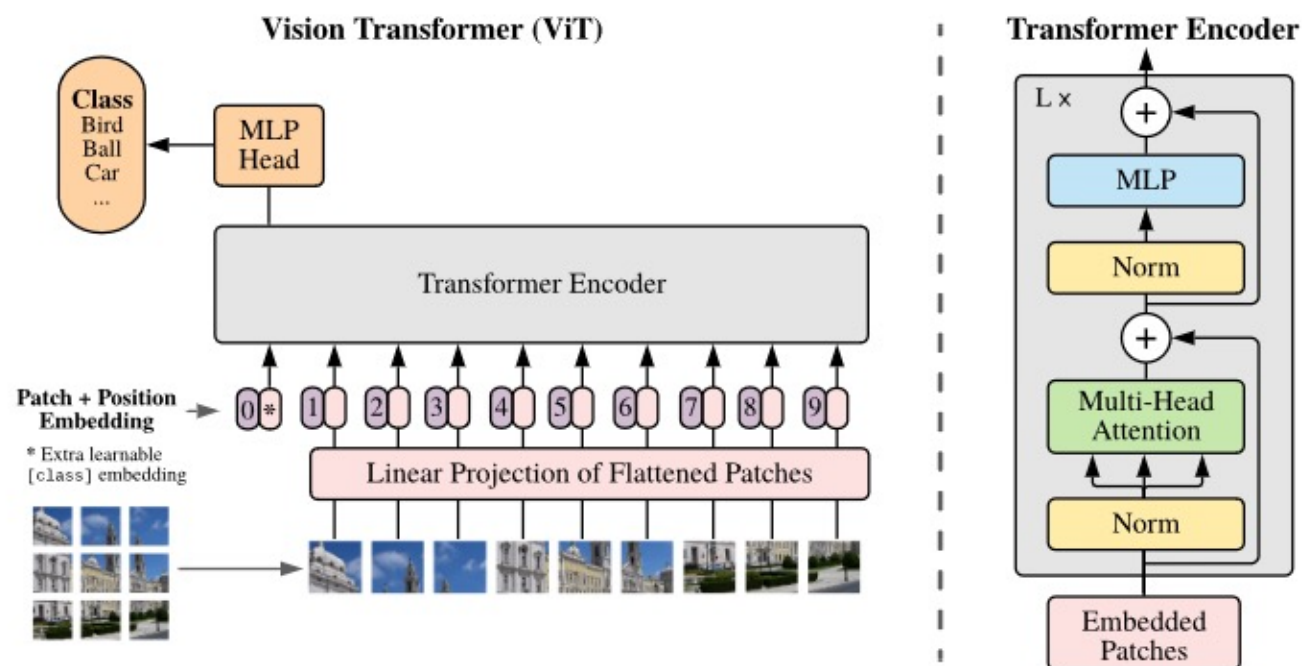


$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$\text{MultiHead}(Q, K, V) = \text{concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_H)W^o$$

例：ViT (Visual Transformer)

网络结构



模型训练

- 损失函数：交叉熵损失（图片分类）

➤ 第13章人工神经网络

■ 神经网络模型的训练：梯度下降

- **梯度消失**与梯度爆炸
- 梯度计算：反向传播、批处理梯度下降、随机梯度下降、小批量梯度下降
- 自适应的梯度下降：动量法、自适应学习率调整
- 参数初始化

小的随机数：方差的确定

预训练模型

BN

■ 神经网络抗过拟合

- 及早停止
- 正则
- 数据增广
- **Dropout**