

AI2611 CheatSheet

Haoyu Zhen

2022 年 6 月 20 日

Preface

Introduction

- My Machine Learning Lecture Notes (更丰富的版本, 适合学习这门课):
[Self-Learning/CS229/MyNotes/ML.pdf](#)
- L^AT_EX code of the CheatSheet is available at:
[SJTU-Course-Stack/AI2611/CheatSheet](#)
- CheatSheet 为 2022 版 (非常少较草率, 当然有部分超纲, 不适合用于学习)。

Acknowledgement

- CS229, Stanford University. (Foundation, Kernel Method, GMM)
- CS231n, Stanford University. (Deep Learning)
- CS189, UC Berkeley. (Random Forest)
- Understanding Machine Learning, Cambridge University. (Linear Regression, KNN, SVM, K-means, PCA)
- AI2611, Shanghai Jiao Tong University. (中文部分, Spectral, Dimensionality Reduction)

§ Foundation

监督学习、无监督学习（聚类、降维）
独立同分布；最小化范化误差

Actual class	Predicted class	
	TP True Positives	FN False Negatives Type II error
+		
-	FP False Positives Type I error	TN True Negatives

Metric	Formula	Interpretation
准确率 (Accuracy)	$\frac{TP + TN}{TP + TN + FP + FN}$	Overall performance of model
查准率 (Precision)	$\frac{TP}{TP + FP}$	How accurate the positive predictions are
通过率、查全率 (Recall Sensitivity)	$\frac{TP}{TP + FN}$	Coverage of actual positive sample
假阳率 (FPR)	$\frac{FP}{TN + FP}$	
F1 score	$\frac{2TP}{2TP + FP + FN}$	Hybrid metric useful for unbalanced classes
F_β score	$\frac{(1 + \beta^2) \times \text{Pre} \times \text{Rec}}{\beta^2 \times \text{Pre} + \text{Rec}}$	Precision and Recall

在不同的阈值下可以得到不同的 TPR 和 FPR 值，将它们在图中绘制出来，并依次连接起来就得到了 ROC 曲线，阈值取值越多，ROC 曲线越平滑。AUC：ROC 曲线下的面积

模型选择：留出法 (hold-out)，保持数据分布的一致性，多次重复划分取平均值)、交叉验证法 (留一法，10-fold，更接近期望评估的模型，但计算量巨大；测试集只包含一个数据，无法分层采样，测试误差差别较大)；自助法 (bootstrap，有放回采样，训练集中数据存在重复，适合小规模数据集)

误差-方差分解

$$E(f; D) = \text{bias}^2 + \text{variance} + \text{error}^2 \\ = (\bar{f}(x) - y)^2 + E_D[f(x; D) - \bar{f}(x)] + E_D[(y_D - y)^2]$$

§ Linear Regression

Loss function: $\mathcal{L}_S(h) = \frac{1}{m} \sum_{i=1}^m (h(\mathbf{x}) - y_i)^2$. Solve $A\mathbf{w} = \mathbf{b}$
where $A \stackrel{\text{def}}{=} \sum \mathbf{x}_i \mathbf{x}_i^T = \mathbf{X} \mathbf{X}^T$ and $\mathbf{b} \stackrel{\text{def}}{=} \sum y_i \mathbf{x}_i = \mathbf{X}^T \mathbf{y}$.

Theorem 1 $\omega = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$. 优点: single-shot 算法，易于实现；缺点：伪逆计算量大可能导致数值不稳定 (奇异矩阵)。

Theorem 2 Eigenvalue decomposition: $A = VD^+V^T$ where D is a diagonal matrix and V is an orthonormal matrix. Define D^+ to be the diagonal matrix: $D_{i,i}^+ = 0$ if $D_{i,i} = 0$ otherwise $D_{i,i}^+ = 1/D_{i,i}$. Then, $A\hat{\mathbf{w}} = \mathbf{b}$ where $\hat{\mathbf{w}} = VD^+V^T\mathbf{b}$.

Remark 1 Gradient Descent. 优点: 收敛快，易于实现；缺点：批量更新，可伸缩性问题。

概率解释: $y|x; \theta \sim \mathcal{N}(0, \sigma)$. Loss function 为最大似然的结果，分类器为 $E[y|x]$ 。

Ridge (岭) Regression

$R_{\text{regularization}}(\mathbf{w}) = \lambda \|\mathbf{w}\|^2$ and $\mathbf{w} = (2\lambda mI + A)^{-1}$. λ 适当增大可以减少方差，但会提高误差。

Lasso (索套) Regression $R(\mathbf{w}) = \lambda \|\mathbf{w}\|_1$. SPARSE

KNN 超参数: k 和 $d(x_1, x_2)$. 计算量大，The “Curse of Dimensionality”， $m \geq (4c\sqrt{d}/\varepsilon)^{d+1}$. 低维度 + 边界非线性 + 密度高时使用。

贝叶斯 先验: $\Pr(\omega_j)$, 后验: $\Pr(\omega_j|x)$, Likelihood: $\Pr(x|\omega_j)$ and Post = likely \propto Prior / $\Pr(x)$ where ω_j 表示类别， x 表示数据 (特征)。

最小化采取 α_i 行动的风险: $R(\alpha_i|x) = \sum_{j=1}^c \lambda(\alpha_i|\omega_j) \Pr(\omega_j|x)$ ，其中 λ 表示在自然状态为 ω_j 的情况下因采取行动 α_i 而产生的损失。If $R(\alpha_1|x) < R(\alpha_2|x)$, then we adopt α_1 (ω_1). Usually,

$$\lambda(\alpha_i, \omega_j) = 1 - \delta_{ij} \quad \text{and} \quad R(\alpha_i|x) = 1 - \Pr(\omega_i|x)$$

参数估计

$$\hat{\theta} = \arg \max_{\theta} \mathcal{L}(\theta) = \sum_{k=1}^n \log \Pr(x_k|\theta).$$

§ Random Forest

Definition 1 (Entropy)

$$H(Y) = -\sum_k \Pr(Y = k) \log \Pr(Y = k) \\ H(Y|X_j) = \Pr(X_j = 1) H(Y|X_j = 1) \\ + \Pr(X_j = 0) H(Y|X_j = 0)$$

Mutual information between X_j and Y .

$$\max I(X_j; Y) \triangleq H(Y) - H(Y|X_j)$$

Gini impurity/index: $G(Y) = 1 - \sum_k \Pr^2(Y = k)$

Random Forest

Ensemble method+randomized+reduce correlation

- bagging (bootstrap aggregating):** sample some data points uniformly with replacement, and use these as the training set.
- feature randomization:** sample some number $k < d$ of features as candidates to be considered for this split.

§ LDA

类间 $d = |w^T \mu_1 - w^T \mu_2|$, 类内 $d_j = \sum_{i \in C_j} |w^T x_i - w^T \mu_j|$

从而 $d_w = d_1^2 + d_2^2 = w^T \Sigma w$.

Solve: $\max(d^2/d_w) = \max(w^T S_b w / w^T S_w w) \triangleq \lambda(w)$.

Theorem 3 $S_w^{-1} S_b w = \lambda w$, w 为最大特征值所对应的特征向量。

Remark 2 多类问题最多可以降至 $M - 1$ 维。

Calinski-Harabasz index:

$$S_b = \sum_{j=1}^M N_j (\mu_j - \mu)(\mu_j - \mu)^T \\ S_w = \sum_{j=1}^M \sum_{i \in C_j} (x_i - \mu_j)(x_i - \mu_j)^T$$

§ Kernel Method $K(x, z) \triangleq (\phi(x), \phi(z))$

Remark 3 Kernel is a corresponding to the feature map ϕ as a function that maps $\mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$.

Definition 2 (Gaussian kernel)

$$K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right).$$

The gaussian kernel is corresponding to an infinite dimensional feature mapping ϕ . Also, ϕ lives in Hilbert space.

§ SVM

$$\arg \max_{(w, b): \|w\|=1} \min_{i \in [m]} |w^T x_i + b| \quad \text{s.t. } \forall i, y^i (w^T x^i + b) \geq 1. \\ \arg \max_{(w, b): \|w\|=1} \min_{i \in [m]} y^i (w^T x^i + b) \quad (1)$$

$$(w_0, b_0) = \arg \min_{(w, b)} \frac{1}{2} \|w\|^2 \quad \text{s.t. } \forall i, y^i (w^T x^i + b) \geq 1.$$

Output: $\hat{w} = w_0/\|w_0\|$, $\hat{b} = b_0/\|w_0\|$

Support Vector: $g(x) = \pm 1$. $\gamma \sim \frac{1}{\|w\|}$

Soft-SVM and Norm Regularization

$$\min_{w, b, \xi} \left(\lambda \|w\|^2 + \frac{1}{m} \sum_{i=1}^m \xi_i \right) \\ \text{s.t. } \forall i, y^i (w^T x^i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

Output: w, b

Definition 3 (hinge loss)

$$l^{\text{hinge}}((w, b), (x, y)) = \max\{0, 1 - y w^T x + b\}.$$

Now we just need to optimize $\lambda \|w\|^2 + \mathcal{L}^{\text{hinge}}(w, b)$. 松弛变量几何意义：对越界数据惩罚力度。

§ K-means 收敛：递减 + 有下界。

前輩 将采样数据从某一分组分类到另一分组，目的是使得损失函数 $\min J = \sum_{i=1}^c J_i = \sum_{i=1}^c \sum_{x \in H_i} \|x - \mu_i\|^2$. \hat{x} 从类别 i 移至 j ，更新公式为: $m_j^* = m_j + \frac{\hat{x} - \mu_j}{n_j + 1}$, $J_j^* = J_j + \frac{n_j}{n_j + 1} \|\hat{x} - \mu_j\|^2$ and $J_i^* = J_i - \frac{n_i}{n_i - 1} \|\hat{x} - \mu_i\|^2$. Transfer \hat{x} to H_k whose $J_k^* - J_l$ is smallest.

The Choice of k

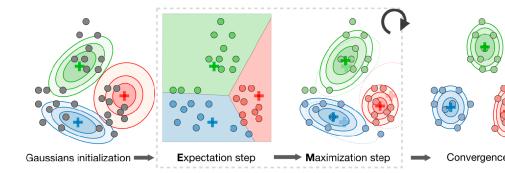
- The Elbow Method: Calculate the Within Cluster Sum of Squared Errors for different values of k , and choose the k for which WSS becomes first starts to diminish.
- The Silhouette value: $a(i) = \frac{1}{|C_I|-1} \sum_{j \in C_I, j \neq i} d(i, j)$, $b(i) = \min_{j \neq i} \frac{1}{|C_J|} \sum_{j \in C_J} d(i, j)$ and $s(i) = \frac{b(i)}{a(i)}$.

缺点：样本数据发生很小的扰动，那么样本的分类结果容易发生明显的改变。
Linkage-Based Clustering Algorithms aka Agglomerative Clustering which is trivial. Stopping criteria: numbers of clusters or current distance.

GMM $x|z \sim \mathcal{N}(\mu, \Sigma)$ where $z \sim \text{Mul}(\phi)$ is the latent variable

- E-step: Evaluate the posterior $\Pr: Q_i(z^i) = P(z^i|x^i; \theta)$.
- M-step: Use the posterior $\Pr Q_i(z^i)$ as cluster specific weights on data points x^i to separately re-estimate each cluster model:

$$\theta_i = \arg \max_{\theta} \sum_i \int_{z^i} Q_i(z^i) \log \left(\frac{P(x^i, z^i; \theta)}{Q_i(z^i)} \right) dz^i$$



Likelihood:

$$\mathcal{L}(\phi, \mu, \Sigma) = \sum_{i=1}^n \log \sum_{z^i=1}^k \Pr(x^i|z^i; \mu, \Sigma) \Pr(z^i|\phi)$$

Spectral 对数据有很好的表征，非凸数据；但是拓展性不好。

Similarity: $W_{ij} = s(x_i, x_j) = \exp(-\|x_i - x_j\|^2/2\sigma^2)$

Graph Constructin: ε -neighborhood, fully connected and KNN.

Definition 4 Some useful definitions:

$$d_i = \sum_{j \in V} W_{ij} \\ \text{cut}(A, B) = \sum_{i \in A, j \in B} W_{ij} \\ \text{cut}(A_1, A_2, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k \text{cut}(A_i, \bar{A}_i) \\ \text{RatioCut}(A_1, A_2, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k \text{cut}(A_i, \bar{A}_i)/|\text{vol}(A_i)| \\ \text{Ncut}(A_1, A_2, \dots, A_k) = \frac{1}{2} \sum_{i=1}^k \text{cut}(A_i, \bar{A}_i)/\text{vol}(A_i) \\ \text{Degree of the subgraph A: } \text{vol}(A) = d_A = \sum_{i, j \in A} W_{ij} \\ \text{Laplacian: } L = D - W \text{ where } W = \text{diag}(\sum_{(i,j) \in E} W_{ij})$$

Theorem 4

$\text{cut}(A, \bar{A}) = x^T L x$ and $\text{cut}(A_1, \dots, A_k) = \text{trace}(X^T L X)$

Proof: $LHS = \sum_{i \in A} d_i - \sum_{i, j \in A} W_{ij}$.

Theorem 5 Relaxation: $x \in \{0, 1\}^{|V|} \rightarrow \mathbb{R}^{|V|}$.

$\min \text{cut}(A, \bar{A}) \iff \min x^T L x = (x^T L x)/(x^T x) \iff x$ 为 L 的最小非零特征值所对应的特征向量 (Rayleigh quotient Theorem)。

Definition 5 Normalized Spectral Clustering:

- 对拉普拉斯矩阵进行标准化操作: $L \leftarrow D^{-0.5} L D^{-0.5}$
- 计算标准化操作后拉普拉斯矩阵最小的 k_1 个特征值对应的特征向量 F
- 将 F 组成的矩阵按行标准化，组成 $n \times k_1$ 的特征矩阵 P 。
- 对 P 中的每一行作为一个 k_1 维的样本，用聚类方法进行聚类。

PCA Compressing matrix $W \in \mathbb{R}^{n,d}$ and recovering matrix $U \in \mathbb{R}^{d,n}: \arg \min_{W, U} \sum_{i=1}^n \|x_i - UWx_i\|^2$

Lemma 1 Let (U, W) be a solution of Equation above. Then $U^T U = I$ and $W = U^T$. (The columns of U are orthonormal.)

By the fact that

$$\|x - UU^T x\|^2 = \|x\|^2 - \text{trace}(U^T x x^T U)$$

We could rewrite the Equation as follows:

$$\arg \max_{U \in \mathbb{R}^{d,n}: U^T U = I} \text{trace}\left[U^T \left(\sum_{i=1}^n x_i x_i^T\right) U\right]$$

Theorem 6 Let x_1, \dots, x_m be arbitrary vectors in \mathbb{R}^d , let $A = \sum_{i=1}^m x_i x_i^T$, and let u_1, \dots, u_n be n eigenvectors of the matrix A corresponding to the largest n eigenvalues of A . Then, the solution to the PCA optimization problem given in Equation is to set U to be the matrix whose columns are u_1, \dots, u_n and to set $W = U^T$. (More Intuition: MathOverflow)

Proof 1 Let $V D V^T$ be the spectral decomposition of A (suppose that $D_{1,1} \geq \dots \geq D_{d,d}$) and let $B = V^T U$. We have

$$\text{trace}(U^T A U) = \text{trace}(B^T D B) = \sum_{j=1}^d D_{j,j} \sum_{i=1}^n B_{j,i}^2 \\ \leq \max_{\beta \in [0, 1]^d: \|\beta\| \leq n} \sum_{j=1}^d D_{j,j} \beta_j = \sum_{j=1}^n D_{j,j}$$

Nota Bene: $B^T B = I$ which entails $\sum_{j=1}^n \sum_{i=1}^m B_{j,i}^2 = n$.

Dimensionality Reduction

PCA 拟合了训练数据的长轴短轴，使得映射后得到的低维度向量分布散射最大
MDS 找到映射方向使得在低维空间中高维度样本间距离不变

$$\min_{\mathbf{x} \in \mathbb{R}^d} \sum_{i < j} (\|\hat{x}_i - \hat{x}_j\| - d_{ij})^2$$

ISOMAP Geodesic 距离 ($d_{ij} \leftarrow$ shortest path) 能反映该数据的真正低维流形结构，保留数据集的本征几何特征。

Locally Linear Embedding 从局部的线性结构关系，恢复全局的非线性流形。假设: $\hat{x}_i = \sum_j W_{ij} x_j$ and $\sum_j W_{ij} = 1$.

$$\min_{\mathbf{x} \in \mathbb{R}^d} \left\| \mathbf{x} - \sum_j W_{ij} x_j \right\|^2 \text{ s.t. } W_{ij} = 0 \text{ if } x_j \in \mathcal{N}(x_i)$$

ISOMAP VS LLE 都保留了邻接的几何结构；都没有显式的映射函数，故使用比较麻烦；LLE 需要更多的训练数据；ISOMAP 计算效率高，实用性更高

§ Deep Learning Trivial.

CNN

$$N = \frac{W - F + 2P}{S} + 1.$$

RNN

$$h_{t+1} = \text{Tanh}(W_h h_t + W_x x_{t+1}).$$

Attention

$$y = \text{SoftMax}\left(\frac{Q K^T}{\sqrt{D}}\right) V.$$

Weight Initialization

$$w \leftarrow \sqrt{2/n} \times \text{RandN}(0, 1)$$

Activate Function

$$\text{Sigmoid}(z) = \frac{1}{1 + \exp(-z)} \quad \text{and} \quad \text{Tanh}(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

Batch Normalization (BP for BN is challenging)

Update:

$$\mu = \alpha \mu + (1 - \alpha) \hat{\mu} \quad \text{and} \quad \sigma^2 = \alpha \sigma^2 + (1 - \alpha) \hat{\sigma}^2.$$

Forward and Eval:

$$x_i \leftarrow \gamma \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta$$

Momentum

$$v \leftarrow \mu v - \alpha dx \quad \text{and} \quad x \leftarrow x + v$$

Adam

$$m \leftarrow \beta_1 m + (1 - \beta_1) dx \quad \text{and} \quad m_t \leftarrow m / (1 - \beta_1^t) \\ v \leftarrow \beta_2 v + (1 - \beta_2) dx^2 \quad \text{and} \quad v_t \leftarrow v / (1 - \beta_2^t)$$

$$x \leftarrow x - \frac{\alpha m_t}{\sqrt{v_t + \epsilon}}$$