机器学习与深度学习习题集（上）

本文是 SIGAI 公众号文章作者编写的机器学习和深度学习习题集（上），是《机器学习-原理、算法与应用》一书的配套产品。此习题集课用于高校的机器学习与深度学习教学，以及在职人员面试准备时使用。为了帮助高校更好的教学，我们将会对习题集进行扩充与优化，并免费提供给高校教师使用。对此感兴趣的在校教师和学生可以通过向 SIGAI 微信公众号发消息获取。习题集的下半部分、所有题目的答案将在后续的公众号文章中持续给出。

第 2 章 数学知识

包括微积分，线性代数与矩阵论，概率论与信息论，最优化方法 4 部分。

1.计算下面函数的一阶导数和二阶导数：

$$f(x) = x\ln x - \frac{1+\exp(2x)}{1-\exp(2x)}$$

2.计算下面两个向量的内积：

$$\mathbf{x} = [1 \quad 2 \quad 3]$$
$$\mathbf{y} = [-1 \quad 5 \quad 10]$$

3.计算下面向量的 1 范数和 2 范数：

$$\mathbf{x} = [1 \quad -2 \quad 3]$$

4.计算下面两个矩阵的乘积：

$$\mathbf{A}=\begin{bmatrix}1 & 2 & 1\\ 0 & 1 & 1\end{bmatrix}$$

$$B=\begin{bmatrix}1 & 2 & 1 & 1\\ 5 & 4 & 0 & 1\\ 7 & 6 & 1 & 0\end{bmatrix}$$

5.计算下面多元函数的偏导数：

$$f\left(x_1,x_2,x_3\right)=\ln\left(1+\exp\left(-2x_1+3x_2-4x_3\right)\right)$$

6.计算下面多元函数的梯度：

$$f\left(x_1,x_2,x_3\right)=\ln\left(1+\exp\left(-2x_1^2+3x_2^3-4x_3\right)\right)$$

7.计算下面多元函数的雅克比矩阵：

$$f\left(x_1,x_2,x_3\right)=x_1^2-\ln x_2+\exp\left(x_1x_3\right)$$

8.计算下面多元函数的 Hessian 矩阵：

$$f\left(x_1,x_2,x_3\right)=x_1^2-\ln x_2+\exp\left(x_1x_3\right)$$

9.计算下面函数的所有极值点，并指明是极大值还是极小值：

$$f\left(x\right)=x^3+2x^2-5x+10$$

10.推导多元函数梯度下降法的迭代公式。

11.梯度下降法为什么要在迭代公式中使用步长系数？

12.梯度下降法如何判断是否收敛？

13.推导多元函数牛顿法的迭代公式。

14.如果步长系数充分小，牛顿法在每次迭代时能保证函数值下降吗？

15.梯度下降法和牛顿法能保证找到函数的极小值点吗，为什么？

16.解释一元函数极值判别法则。

17.解释多元函数极值判别法则。

18.什么是鞍点？

19.解释什么是局部极小值，什么是全局极小值。

20.用拉格朗日乘数法求解如下极值问题

$$\min f\left(x_1, x_2\right) = x_1^2 - 4x_1x_2 + x_2^2$$
$$x_1 + x_2 + x_3 = 1$$

21.什么是凸集？

22.什么是凸函数，如何判断一个一元函数是不是凸函数，如何判断一个多元函数是不是凸函数？

22.什么是凸优化？

23.证明凸优化问题的局部最优解一定是全局最优解。

24.对于如下最优化问题：

$$\min f\left(x_1, x_2\right) = x_1^2 - 4x_1x_2 + x_2^2$$
$$x_1 + x_2 + x_3 = 1$$
$$x_1 - x_2 + x_3 > 0$$

构造广义拉格朗日乘子函数，将该问题转化为对偶问题。

25.一维正态分布的概率密度函数为

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right)$$

给定一组样本 $x_1,...,x_l$ 。用最大似然估计求解正态分布的均值和方差。

26.如何判断一个矩阵是否为正定矩阵？

27. 解释最速下降法的原理。

28.解释坐标下降法的原理。

29.一维正态分布的概率密度函数为

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(x-\mu\right)^2}{2\sigma^2}\right)$$

按照定义计算其数学期望与方差。

30.两个离散型概率分布的 KL 散度定义为：

$$D_{\mathrm{KL}}\left(P\|Q\right)=\sum_{x}P\left(x\right)\ln\frac{P\left(x\right)}{Q\left(x\right)}$$

利用下面的不等式，当$x>0$时：

$$\ln x\le x-1$$

证明 KL 散度非负，即

$$D_{\mathrm{KL}}\left(p\|q\right)\ge 0$$

31.对于离散型概率分布，证明当其为均匀分布时熵有最大值。

32.对于连续型概率分布，已知其数学期望为μ，方差为σ^2。用变分法证明当此分布为正态分布时熵有最大值。

33.对于两个离散型概率分布，证明当二者相等时交叉熵有极小值。

34.为什么在实际的机器学习应用中经常假设样本数据服从正态分布？

35.什么是随机事件独立，什么是随机向量独立？

36.什么是弱对偶？什么是强对偶？

37.证明弱对偶定理。

38.简述 Slater 条件。

39.简述 KKT 条件。

40.解释蒙特卡洛算法的原理。为什么蒙特卡洛算法能够收敛？

41.解释熵概念。

第 3 章 基本概念

1.名词解释：有监督学习，无监督学习，半监督学习。

2.列举常见的有监督学习算法。

3.列举常见的无监督学习算法。

4.简述强化学习的原理。

5.什么是生成模型？什么是判别模型？

6.概率模型一定是生成模型吗？

7.不定项选择。下面那些算法是生成模型？_____哪些算法是判别模型？_____
A.决策树 B.贝叶斯分类器 C.全连接神经网络 D.支持向量机 E. logistic 回归
F. AdaBoost 算法 G.隐马尔可夫模型 H.条件随机场 I.受限玻尔兹曼机

8.如何判断是否发生过拟合？

9.发生过拟合的原因有哪些，应该怎么解决？

10.列举常见的正则化方法。

11.解释 ROC 曲线的原理。

12.解释精度，召回率，F1 值的定义。

13.解释交叉验证的原理。

14.什么是过拟合，什么是欠拟合？

15.什么是没有免费午餐定理？

16.简述奥卡姆剃刀原理。

17.推导偏差-方差分解公式。

18.证明如果采用均方误差函数，线性回归的优化问题是凸优化问题。

19.推导线性回归的梯度下降迭代公式。

20.解释混淆矩阵的概念。

21.解释岭回归的原理。

22.解释 LASSO 回归的原理。

第 4 章 贝叶斯分类器

1.什么是先验概率，什么是后验概率？

2.推导朴素贝叶斯分类器的预测函数。

3.什么是拉普拉斯光滑？

4.推导正态贝叶斯分类器的预测函数。

5.贝叶斯分类器是生成模型还是判别模型？

第 5 章 决策树

1.什么是预剪枝，什么是后剪枝？

2.什么是属性缺失问题？

3.对于属性缺失问题，在训练时如何生成替代分裂规则？

4.列举分类问题的分裂评价指标。

5.证明当各个类出现的概率相等时，Gini 不纯度有极大值；当样本全部属于某一类时，Gini 不纯度有极小值。

6.ID3 用什么指标作为分裂的评价指标？

7.C4.5 用什么指标作为分裂的评价指标？

8.解释决策树训练时寻找最佳分裂的原理。

9.对于分类问题，叶子节点的值如何设定？对于回归问题，决策树叶子节点的值如何设定？

10.决策树如何计算特征的重要性？

11.CART 对分类问题和回归问题分别使用什么作为分裂评价指标？

第 6 章 k 近邻算法与距离度量学习

1.简述 k 近邻算法的预测算法的原理。

2.简述 k 的取值对 k 近邻算法的影响。

3.距离函数需要满足哪些数学条件？

4.列举常见的距离函数。

5.解释距离度量学习的原理。

6.解释 LMNN 算法的原理。

7.解释 ITML 算法的原理。

8.解释 NCA 算法的原理。

第 7 章 数据降维

1.使用数据降维算法的目的是什么？

2.列举常见的数据降维算法。

3.常见的降维算法中，哪些是监督降维，哪些是无监督降维？

4.什么是流形？

5.根据最小化重构误差准则推导 PCA 投影矩阵的计算公式。

6.解释 PCA 降维算法的流程。

7.解释 PCA 重构算法的流程。

8.解释 LLE 的原理。

9.名词解释：图的拉普拉斯矩阵。

10.解释 t-SNE 的原理。

11.解释 KPCA 的原理。

12.证明图的拉普拉斯矩阵半正定。

13.解释拉普拉斯特征映射的原理。

14.解释等距映射的与原理。

15.PCA 是有监督学习还是无监督学习？

第 8 章 线性判别分析

1.解释 LDA 的原理。

2.推导多类和高维时 LDA 的投影矩阵计算公式。

3.解释 LDA 降维算法的流程。

4.解释 LDA 重构算法的流程。

5.LDA 是有监督学习还是无监督学习？

第 9 章 人工神经网络

1.神经网络为什么需要激活函数？

2.推导 sigmoid 函数的导数计算公式。

3.激活函数需要满足什么数学条件？

4.为什么激活函数只要求几乎处处可导而不需要在所有点处可导？

5.什么是梯度消失问题，为什么会出现梯度消失问题？

6.如果特征向量中有类别型特征，使用神经网络时应该如何处理？

7.对于多分类问题，神经网络的输出值应该如何设计？

8.神经网络参数的初始值如何设定？

9.如果采用欧氏距离损失函数，推导输出层的梯度值。推导隐含层参数梯度的计算公式。

10.如果采用 softmax+交叉熵的方案，推导损失函数对 softmax 输入变量的梯度值。

11.解释动量项的原理。

12.列举神经网络的正则化技术。

13.推导 ReLU 函数导数计算公式。

第 10 章 支持向量机

1.推导线性可分时 SVM 的原问题：

$$\min \frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w}$$
$$y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i+b\right)\geq 1$$

2.证明线性可分时 SVM 的原问题是凸优化问题且 Slater 条件成立：

$$\min \frac{1}{2}\mathrm{w}^{\mathrm{T}}\mathrm{w}$$
$$y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i+b\right)\geq 1$$

3.推导线性可分时 SVM 的对偶问题：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \mathrm{x}_i^{\mathrm{T}}\mathrm{x}_j-\sum_{i=1}^{l}\alpha_i$$
$$\alpha_i\geq 0, i=1,...,l$$
$$\sum_{i=1}^{l}\alpha_i y_i=0$$

4.证明加入松弛变量和惩罚因子之后，SVM 的原问题是凸优化问题且 Slater 条件成立：

$$\min \ \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}+C\sum_{i=1}^{l}\xi_i$$
$$y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b)\geq 1-\xi_i$$
$$\xi_i\geq 0, i=1,...,l$$

5.推导线性不可分时 SVM 的对偶问题：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \mathrm{x}_i^{\mathrm{T}}\mathrm{x}_j-\sum_{i=1}^{l}\alpha_i$$
$$0\leq\alpha_i\leq C$$
$$\sum_{j=1}^{l}\alpha_j y_j=0$$

6.证明线性不可分时 SVM 的对偶问题是凸优化问题：

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j \mathrm{x}_i^{\mathrm{T}}\mathrm{x}_j-\sum_{i=1}^{l}\alpha_i$$
$$0\leq\alpha_i\leq C$$
$$\sum_{j=1}^{l}\alpha_j y_j=0$$

7.用 KKT 条件证明 SVM 所有样本满足如下条件：

$$\alpha_i = 0 \Rightarrow y_i(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i + b) \geq 1$$
$$0 < \alpha_i < C \Rightarrow y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i + b\right) = 1$$
$$\alpha_i = C \Rightarrow y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i + b\right) \leq 1$$

8.SVM 预测函数中的b值如何计算？

9.解释核函数的原理，列举常用的核函数。

10.什么样的函数可以作为核函数？

11.解释 SMO 算法的原理。

12.SMO 算法如何挑选子问题的优化变量？

13.证明 SMO 算法中子问题是凸优化问题。

14.证明 SMO 算法能够收敛。

15.SVM 如何解决多分类问题？

第 11 章 线性模型

1.logistic 回归中是否一定要使用 logistic 函数得到概率值？能使用其他函数吗？

2.名称解释：对数似然比。

3.logistic 是线性模型还是非线性模型？

4.logistic 回归是生成模型还是判别模型？

5.如果样本标签值为 0 或 1，推导 logistic 回归的对数似然函数：

$$\ln L(\mathbf{w}) = \sum_{i=1}^{l}\left(y_i \log h(\mathbf{x}_i) + (1-y_i)\log\left(1-h(\mathbf{x}_i)\right)\right)$$

6.logistic 回归中为什么使用交叉熵而不使用欧氏距离作为损失函数？

7.证明 logistic 回归的优化问题是凸优化问题：

$$f(\mathrm{w})=-\sum_{i=1}^{l}\left(y_i\log h(\mathbf{x}_i)+(1-y_i)\log\left(1-h(\mathbf{x}_i)\right)\right)$$

8.推导 logistic 回归的梯度下降迭代公式。

9.如果类别别标签为+1 和-1，推导 logistic 回归的对数似然函数：

$$-\sum_{i=1}^{l}\log\left(1+\exp\left(-y_i\left(\mathrm{w}^{\mathrm{T}}\mathrm{x}_i+b\right)\right)\right)$$

10.写出使用 L1 和 L2 正则化项时 logistic 回归的目标函数。

11.写出 softmax 回归的预测函数。

12.推导 softmax 回归的对数似然函数：

$$\sum_{i=1}^{l}\sum_{j=1}^{k}\left(y_{ij}\ln\frac{\exp\left(\boldsymbol{\theta}_j^{\mathrm{T}}\mathbf{x}_i\right)}{\sum_{t=1}^{k}\exp\left(\boldsymbol{\theta}_t^{\mathrm{T}}\mathbf{x}_i\right)}\right)$$

13.证明 softmax 回归的优化问题是凸优化问题。

14.推导 softmax 回归的梯度计算公式。

15.logistic 回归如何计算特征的重要性？

第 12 章 随机森林

1.解释 Bagging 算法的原理。

2.解释随机森林预测算法对分类问题，回归问题的处理。

3.随机森林如何输出特征的重要性？

4.解释随机森林预测算法的原理。

5.随机森林为什么能够降低方差？

第 13 章 Boosting 算法

1.写出 AdaBoost 算法强分类器的预测公式。

2.写出 AdaBoost 的训练算法。

3.证明强分类器在训练样本集上的错误率上界是每一轮调整样本权重时权重归一化因子的乘积，即下面的不等式成立：

$$p_{error}=\frac{1}{l}\sum_{i=1}^{l}\left[\left[\operatorname{sgn}\left(F\left(\mathrm{x}_i\right)\right)\neq y_i\right]\right]\leq\prod_{t=1}^{T}Z_t$$

4.证明下面的不等式成立：

$$\prod_{t=1}^{T}Z_t=\prod_{t=1}^{T}2\sqrt{e_t\left(1-e_t\right)}=\prod_{t=1}^{T}\sqrt{\left(1-4\gamma_t^2\right)}\leq\exp\left(-2\sum_{t=1}^{T}\gamma_t^2\right)$$

5.简述广义加法模型的原理。

6.离散型 AdaBoost 的损失函数是什么函数？

7.从广义加法模型和指数损失函数推导 AdaBoost 的训练算法。

8.解释实数型 AdaBoost 算法的原理。

9.AdaBoost 算法的弱分类器应该如何选择？

10.简述梯度提升算法的原理。

11.假设使用均方误差函数，梯度提升算法如何解决回归问题？

12.梯度提升算法如何解决二分类问题？

13.对于多分类问题，梯度提升算法的预测函数是 $F_k\left(\mathbf{x}\right)$。样本属于每个类的概率为：

$$p_k\left(\mathbf{x}\right)=\frac{\exp\left(F_k\left(\mathbf{x}\right)\right)}{\sum_{l=1}^{K}\exp\left(F_l\left(\mathbf{x}\right)\right)}$$

如果加上限制条件：

$$\sum_{l=1}^{K}F_l\left(\mathbf{x}\right)=0$$

证明如下结论成立：

$$F_k\left(\mathbf{x}\right)=\ln p_k\left(\mathbf{x}\right)-\frac{1}{K}\sum_{l=1}^{K}\ln p_l\left(\mathbf{x}\right)$$

14.解释 XGBoost 算法的原理。

15.XGBoost 算法为何要泰勒展开到二阶？