1 Classification

1. Derive the dual formulation of the soft support vector machine algorithm.

Solution of T1

软间隔SVM原始问题:
最小化：

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i$$

满足约束：

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0$$

设$\alpha_i \geq 0$,$\mu_i \geq 0$，构造拉格朗日函数：

$$L(w,b,\xi,\alpha,\mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i[y_i(w^Tx_i+b) - 1 + \xi_i] - \sum_{i=1}^{n}\mu_i\xi_i$$

对 w、b、ξ_i 求偏导:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n}\alpha_i y_i x_i = 0 \Rightarrow w = \sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n}\alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \mu_i = 0 \Rightarrow \alpha_i = C - \mu_i$$

代入拉格朗日对偶函数,

$$
\begin{aligned}
\mathcal{G}(\alpha, \mu) = & \min_{w,b,\xi} L(w, b, \xi, \alpha, \mu) \\
= & \frac{1}{2} \| \sum_{i=1}^{n} \alpha_i y_i x_i \|^2 + C \sum_{i=1}^{n} \xi_i \\
& - \sum_{i=1}^{n} \alpha_i [y_i (\sum_{j=1}^{n} \alpha_j y_j x_j^T x_i + b) - 1 + \xi_i] - \sum_{i=1}^{n} \mu_i \xi_i \\
= & \frac{1}{2} \left(\sum_{i=1}^{n} \alpha_i y_i x_i \right)^T \cdot \left(\sum_{j=1}^{n} \alpha_j y_j x_j \right) \\
& - \sum_{i=1}^{n} \alpha_i [y_i (\sum_{j=1}^{n} \alpha_j y_j x_j^T x_i + b) - 1] \\
= & \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j \\
& - b \sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i \\
= & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j
\end{aligned}
$$

从而得到对偶问题：

$$
\max_{\alpha,\mu} \mathcal{G}(\alpha, \mu) = \max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j
$$

约束条件为：

$$
\sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, n
$$

2. Derive the Discriminant function of the Quadratic Discriminant Analysis.

Solution of T2

QDA假设对于每一个类别k，其特征向量x服从多元高斯分布：

$$
p(x|Y=k) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right)
$$

记$\pi_k = P(Y = k)$，由Bayes，后验概率为：

$$
P(Y = k|x) = \frac{\pi_k p(x|Y=k)}{\sum_{l=1}^{K} \pi_l p(x|Y=l)}
$$

决策规则为:

$$
\begin{aligned}
\hat{Y} &= \arg\max_k P(Y = k|x) \\
&= \arg\max_k \frac{\pi_k p(x|Y=k)}{\sum_{l=1}^{K} \pi_l p(x|Y=l)}
\end{aligned}
$$

取对数并忽略常数项：

$$\hat{Y} = \arg\max_k -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log\pi_k$$

从而导出了QDA的判别函数

$$\delta_k(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_k)^\top\Sigma_k^{-1}(\mathbf{x}-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log\pi_k$$

3. Derive the Gaussian process classifier by using the Laplace approximation.

Solution of T3

对分类问题而言，观测目标 $y_i \in \{0,1\}$ 离散，高斯噪声模型 $y_i = f(x_i) + \epsilon_i$ 不再适用。因此引入潜函数 $f(\mathbf{x})$ ，并利用链接函数(通常为sigmoid函数)将其映射到概率。

为潜函数赋予一个高斯过程先验:

$$p(\mathbf{f}|X) = \mathcal{N}(\mathbf{f}|\mathbf{0}, K)$$

其中 $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ 。

给定潜函数 $\mathbf{f}$，$\mathbf{y}$是伯努利随机变量构成的随机向量

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^n p(y_i|f_i) = \prod_{i=1}^n \sigma(f_i)^{y_i}(1-\sigma(f_i))^{1-y_i}$$

要求潜函数的后验分布 $p(\mathbf{f}|X,\mathbf{y})$,根据Bayes: $p(\mathbf{f}|X,\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|X)}{p(\mathbf{y}|X)}$
分母为正则化项。由于 $p(\mathbf{y}|\mathbf{f})$ 非高斯，因此考虑拉普拉斯近似。以下求解近似分布 $q(\mathbf{f}|X,\mathbf{y})$

先求后验分布的众数 $\hat{\mathbf{f}}$。对后验分布取对数，得到：

$$\hat{\mathbf{f}} = \arg\max_{\mathbf{f}}[\log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}|X)]$$

定义 $\Psi(\mathbf{f}) \triangleq \log p(\mathbf{y}|\mathbf{f}) + \log p(\mathbf{f}|X)$
其中

$$\log p(\mathbf{f}|X) = -\frac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{f} - \frac{1}{2}\log|K| - \frac{n}{2}\log 2\pi$$

$$\log p(\mathbf{y}|\mathbf{f}) = \sum_{i=1}^n [y_i\log\sigma(f_i) + (1-y_i)\log(1-\sigma(f_i))]$$

所以，

$$\Psi(\mathbf{f}) = \log p(\mathbf{y}|\mathbf{f}) - \frac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{f} + \text{const.}$$

令 $\nabla\Psi(\mathbf{f}) = \mathbf{0}$，$\nabla\Psi(\mathbf{f}) = \nabla\log p(\mathbf{y}|\mathbf{f}) - K^{-1}\mathbf{f}$
其中， $\nabla\log p(\mathbf{y}|\mathbf{f})$ 的第i个分量为：

$$\frac{\partial}{\partial f_i}\log p(\mathbf{y}|\mathbf{f}) = y_i - \sigma(f_i)$$

因此，

$$\nabla\Psi(\mathbf{f}) = (\mathbf{y} - \boldsymbol{\sigma}) - K^{-1}\mathbf{f}$$

其中 $\boldsymbol{\sigma} = [\sigma(f_1), \ldots, \sigma(f_n)]^T$。从而可以反解出众数 $\hat{\mathbf{f}}$。

以下计算众数处的Hessian矩阵以进行Laplace近似。

$$\nabla\nabla\Psi(\mathbf{f}) = \nabla\nabla \log p(\mathbf{y}|\mathbf{f}) - K^{-1}$$

其中 $\nabla\nabla \log p(\mathbf{y}|\mathbf{f})$ 是一个对角矩阵，其对角线元素为:

$$\frac{\partial^2}{\partial f_i^2} \log p(y_i|f_i) = \frac{\partial}{\partial f_i}(y_i - \sigma(f_i)) = -\sigma(f_i)(1 - \sigma(f_i))$$

所以，在众数 $\hat{\mathbf{f}}$ 处，Hessian矩阵为：

$$H = \nabla\nabla\Psi(\hat{\mathbf{f}}) = -\hat{W} - K^{-1}$$

其中 $\hat{W}$ 是对角阵，$\hat{W}_{ii} = \sigma(\hat{f}_i)(1 - \sigma(\hat{f}_i))$。从而近似分布 $q(\mathbf{f}|X, \mathbf{y}) = \mathcal{N}(\mathbf{f}|\hat{\mathbf{f}}, \Sigma)$，其中协方差矩阵 $\Sigma = (-H)^{-1} = (\hat{W} + K^{-1})^{-1}$。
对于预测点 $\mathbf{x}_*$，下求其分类标签 y_*。

由高斯过程的性质，$\mathbf{f}$ 和 f_* 的联合先验是高斯分布：

$$\begin{bmatrix} \mathbf{f} \\ f_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K & \mathbf{k}_* \\ \mathbf{k}_*^\top & k_{**} \end{bmatrix}\right)$$

其中 $\mathbf{k}_* = [k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_n, \mathbf{x}_*)]^\top$，$k_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$。

则 f_* 的预测分布为：

$$q(f_*|X, \mathbf{y}, \mathbf{x}_*) = \int p(f_*|X, \mathbf{x}_*, \mathbf{f}) q(\mathbf{f}|X, \mathbf{y}) d\mathbf{f} = \mathcal{N}(f_*|\mu_*, \sigma_*^2)$$

其中 $\mu_* = \mathbf{k}_*^\top K^{-1}\hat{\mathbf{f}}, \sigma_*^2 = k_{**} - \mathbf{k}_*^\top (K + \hat{W}^{-1})^{-1}\mathbf{k}_*$

预测概率如下(各项参数见上方表达式):

$$P(y_* = 1|X, \mathbf{y}, \mathbf{x}_*) = \int \sigma(f_*) q(f_*|X, \mathbf{y}, \mathbf{x}_*) df_*$$

4. For the following data, build a tree by using: a) Gini index, b) the depth of tree is two. The target variable is **Loan Approved**.

Solution of T4

首先计算每个特征的Gini

1. Age

取中位数37为分割节点，分成 ≤37, >37 两类。
左侧有10个样本，3个Yes，7个No。
$Gini_L = 1 - ((3/10)^2 + (7/10)^2) = 0.42$
右侧有10个样本，10个Yes，0个No。
$Gini_R = 0$
加权 $Gini_{split} = \frac{10}{20} \times 0.42 + \frac{9}{20} \times 0.1975 = 0.21$

2. Income

取分割点50000，分成 ≤50000, >50000 两类。
左侧有10个样本，3个Yes，7个No。
$Gini_L = 1 - ((3/10)^2 + (7/10)^2) = 0.42$
右侧有10个样本，10个Yes，0个No。

$Gini_R = 0$
加权 $Gini_{split} = \frac{10}{20} \times 0.42 = 0.21$

3. Credit Score

取分割点650，分成 ≤650, >650 两类。
左侧有7个样本，0个Yes，7个No。
$Gini_L = 0$
右侧有13个样本，13个Yes，0个No。
$Gini_R = 0$
加权 $Gini_{split} = 0$

故节点选择Credit Score (≤650 vs >650)，Gini=0，已经达到最小。所以只能将根节点看成第1层，从而得到深度为2的树。

5. For the data set in https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic, try following algorithms learned in the class: a) Discriminate function (mse, ppn, svm); b) Generative method (LDA, QDA, Naive Bayes (Gaussian)); c) logistic regression, neural network; d) KNN, tree method, random forest, e): kernel method (kernel svm, Gaussian process classifier). Report the parameters of your algorithm, and the accuracy on the test set.

Solution of T5

```
[Linear Regression] Accuracy = 0.9766
[Perceptron] Accuracy = 0.9649
  max_iter=1000, tol=1e-3, random_state=42
[Linear SVM] Accuracy = 0.9649
  kernel='linear', C=1.0, random_state=42


[LDA] Accuracy = 0.9649
[QDA] Accuracy = 0.9415
[GaussianNB] Accuracy = 0.9415


[Logistic Regression] Accuracy = 0.9825
  C=1.0, solver='lbfgs', max_iter=1000
[Neural Network MLP] Accuracy = 0.9649
  hidden_layer_sizes=(50,25), activation='relu', solver='adam', max_iter=1000, random_state=42


[KNN] Accuracy = 0.9649
  n_neighbors=5
[Decision Tree] Accuracy = 0.9006
  max_depth=5, random_state=42
[Random Forest] Accuracy = 0.9708
  n_estimators=100, max_depth=5, random_state=42


[RBF SVM] Accuracy = 0.9825
  kernel='rbf', C=1.0, gamma='scale', random_state=42
[Gaussian Process Classifier] Accuracy = 0.9825
  kernel=kernel, random_state=42
```