Machine Learning Assignment 4

Due: May 31, 2024

SVM

1. [35pts] Support Vector Machine

(1) Recall that the soft margin support vector machine solves the problem:

$$min \quad \frac{1}{2} w^\top w + C \sum_i \varepsilon_i$$

$$\text{s.t.} \quad y_i(w^\top x_i + b) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0.$$

a) [10pts] Derive its dual problem using the method of Lagrange multipliers.

b) [10pts] Further simplify the dual problem when at its saddle point to prove

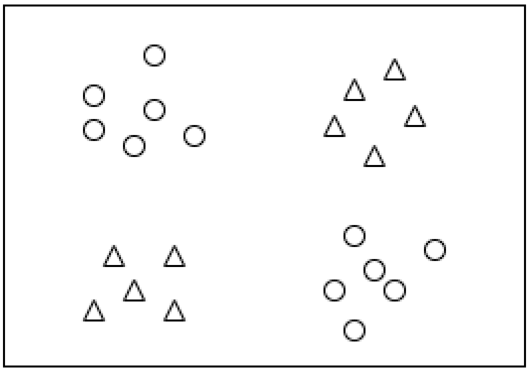
$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \, \alpha_j y_i y_j x_i^\top x_j$$

$$\text{s.t.} \ C \geq \alpha_i \geq 0, \quad \sum_i \alpha_i \, y_i = 0,$$

is equivalent to the primal problem.

(2) [15pts] Given the XOR sample points as below, we train an SVM with a quadratic kernel, i.e. our kernel function is a polynomial kernel of degree 2: $\kappa(x_i, x_j) = (x_i^T x_j)^d, d = 2.$

(a) [5pts] what is the corresponding mapping function $\phi(x)$?



(b) [5pts] Use the following code to generate XOR data, and according to the answer of (a), map the data with $\phi(x)$ to see if it can be linearly separable.

(c) [5pts] Could we get a reasonable model with hard margin (after feature mapping)? If yes, draw the decision boundary in the figure (original feature space), otherwise state reasons.

```
import numpy as np
import matplotlib.pyplot as plt
#创建数据
X_xor = np.random.randn(40,2)
y_xor = np.logical_xor(X_xor[:,0]>0, X_xor[:,1]>0)
y_xor = np.where(y_xor, 1, -1)
#绘制散点图
plt.scatter(x=X_xor[y_xor==1,0]),   #  横坐标
            y=X_xor[y_xor==1,1]),   #  纵坐标
            color=’g’, marker=’x’, label=’1’)
plt.scatter(x=X_xor[y_xor==-1,0]),
            y=X_xor[y_xor==-1,1]),
            color=’b’, marker=’o’, label=’-1’)
plt.legend() #显示图例
plt.show()
```

Solution

2. Kernel Methods [必做题，不提交不批改，可参考教材核对答案]

请给出 kernel PCA 的推导过程。

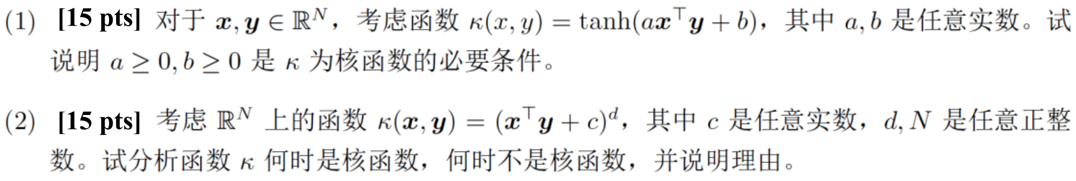
（可中文作答）

Solution

参见课本 10.4 节，或南瓜书对应章节。

3. Kernel Functions [选做题，不提交不批改，后续公布答案]

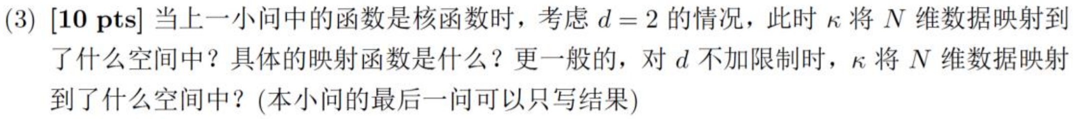
注：其中第(3)小题可以帮助深入理解核函数与特征映射函数之间的关系



(1) **[15 pts]** 对于 $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^N$，考虑函数 $\kappa(x, y) = \tanh(a\boldsymbol{x}^\top \boldsymbol{y} + b)$，其中 a, b 是任意实数。试说明 $a \geq 0, b \geq 0$ 是 κ 为核函数的必要条件。

(2) **[15 pts]** 考虑 $\mathbb{R}^N$ 上的函数 $\kappa(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^\top \boldsymbol{y} + c)^d$，其中 c 是任意实数，d, N 是任意正整数。试分析函数 κ 何时是核函数，何时不是核函数，并说明理由。

说明：该核函数是多项式核的更一般的形式。



(3) **[10 pts]** 当上一小问中的函数是核函数时，考虑 $d = 2$ 的情况，此时 κ 将 N 维数据映射到了什么空间中？具体的映射函数是什么？更一般的，对 d 不加限制时，κ 将 N 维数据映射到了什么空间中？(本小问的最后一问可以只写结果)

（可中文作答）

4. Kernel Methods [选做题，不提交不批改，参考南瓜书核对答案]

推导 kernel LDA

Bayesian Classifiers

1. [30pts] Naïve Bayes Classifier

Suppose you are given the following set of data with four Integer input variables A, B, C, and D, and a single binary label y.

In this task, the value of a variable means how many times it appears in text from the corresponding label (i.e., word frequency, which is a popular representation of text). For example, in the text of x_1 from label +1, word A appears twice, word B appears 4 times, word C appears 10 times and word D appears 3 times.

We are trying to fit a Naïve Bayes Classifier on this dataset.

	A	B	C	D	y
x_1	2	4	10	3	+1
x_2	3	1	4	2	+1
x_3	0	2	0	5	-1
x_4	2	0	4	0	+1
x_5	1	6	6	0	-1
x_6	0	2	1	7	-1
x_7	3	0	0	8	+1
x_8	6	1	2	7	-1

(1) [**20pts**] Calculate the empirical conditional probability of each variable for appearing in texts from each label. To illustrate, for variable A, calculate $p_{A,j} = P(word = A \mid y = j),\ j \in \{-1, +1\}$, and the same for the other variables. Remember to use Laplace smoothing to avoid zero probabilities.

(2) [**10pts**] Give a new sample where $A = 3, B = 2, C = 1, D = 2$. Predict its label. You should write down your calculation in detail. (It is enough to only give the form of a fraction, not necessarily calculated as a decimal, 仅给出分数形式即可，不一定需要计算为小数)

2. [35pts] Gaussian Bayesian Classifiers

Given data set $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, where $y \in Y = \{1,2, \dots, K\}$.

(1) [**5pts**] Please write down the Bayes optimal classifier that minimizes the misclassification error rate.

(2) [**15pts**] Suppose the samples in the k-th class are i.i.d. sampled form normal distribution $\mathcal{N}(\mu_k, \Sigma)$, ($k = 1, 2, \dots, \mathrm{K}$, all classes share the same covariance matrix). Let m_k denote the number of samples in the k-th class, and the prior probability $P(y = k) = \pi_k$. If $x \in R^d \sim \mathcal{N}(\mu, \Sigma)$, then the probability density function is:

$$p(x) = \frac{1}{(2\pi)^{d/2}\, det(\Sigma)^{1/2}} exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

Please write down the corresponding Bayes optimal classifier.

(3) [**15pts**] For binary classification problem, please prove that when samples in each class are i.i.d. sampled from normal distributions which share the same covariance matrix and the two classes have equal prior probabilities $\pi_0 = \pi_1$, LDA (Linear Discriminant Analysis) gives the Bayes optimal classifier.

Hint: The optimal solution of LDA is:

$$w = S_w^{-1}(\mu_0 - \mu_1)$$

where S_w is within-class scatter matrix, $S_w = \Sigma_0 + \Sigma_1$ (Σ_i is the covariance matrix of the *i*-th class).

3. [30pts] MLE and Linear Regression

Sample points come from an unknown distribution, $X_i \sim D$. Labels y_i are the sum of a deterministic function $f(\mathrm{X_i})$ plus random noise: $y_i = f(X_i) + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

For this problem, we will assume that $\varepsilon_i \sim \mathcal{N}(0, \sigma_i^2)$—that is, the variance σ_i^2 of the noise is different for each sample point and we will examine how our loss function changes as a result. We assume that we know the value of each σ_i^2. You are given an $n \times p$ design matrix X, an n-dimensional vector y of labels, such that the label y_i of sample point X_i is generated as described above, and a list of the noise variances σ_i^2.

(1) [**10pts**] Apply MLE to derive the optimization problem that will use the maximum likelihood estimate of the distribution parameter f. (Note: f is a function, but we can still treat it as the parameter of an optimization problem.) Express your Objective function as a summation of loss functions, one per sample point.

(2) [**10pts**] We decide to do linear regression, so we parameterize $f(X_i)$ as $f(X_i) = w \cdot X_i$, where w is a p-dimensional vector of weights. Write an equivalent optimization problem where your optimization variable is w and the cost function is a function of X, y, w, and the variances σ_i^2. Find a way to express your cost function in matrix notation. (Hint: You can define a new matrix.)

(3) [**10pts**] Write the solution to your optimization problem as the solution of a linear system of equations. (Again, in matrix notation.)