CSE/ISYE 6740 Homework 4

Anqi Wu, Fall 2022

Deadline: 12/8 Thursday, 12:30pm ET

- There are 2 sections in gradescope: Homework 4 and Homework 4 Programming. Submit your answers as a PDF file to Homework 4 (including report for programming) and also submit your code in a zip file to Homework 4 Programming.
- All Homeworks are due by the beginning of class. Homework is penalized by 20% for each day that it is late (this applies additively, meaning that no credit is gained after 5 late days).
- We strongly encourage the use of LaTeX for your submission. Unreadable handwriting is subject to zero credit.
- Explicitly mention your collaborators if any.
- Recommended reading: PRML[1] Section 13.2
- Python and Matlab are allowed.
- When submitting to Gradescope, please make sure to mark the page(s) corresponding to each problem/sub-problem. **Note**: This is a large class and Gradescope's assignment segmentation features are essential. RedFailure to follow these instructions may result in parts of your assignment not being graded. We will not entertain regrading requests for failure to follow instructions. Please check this link for additional information on submitting to Gradescope.

1 Kernels [20 points]

TA: Rachmat Subagia

(a) Identify which of the followings is a valid kernel. If it is a kernel, please write your answer explicitly as 'True' and give mathematical proofs. If it is not a kernel, please write your answer explicitly as 'False' and give explanations. [6 pts]

Suppose K_1 and K_2 are valid kernels (symmetric and positive definite) defined on $R^m \times R^m$.

1. $K(u, v) = k_1(u, v) - k_2(u, v)$ for k_1, k_2 valid kernels
2. $K(u, v) = k_1(u, v)k_2(u, v)$ for k_1, k_2 valid kernels
3. $K(u, v) = \exp\left(\gamma\|u - v\|^2\right)$ for some $\gamma > 0$

[1]Christopher M. Bishop, Pattern Recognition and Machine Learning, 2006, Springer.

(b) By considering the determinant of a 2×2 Gram matrix, show that a positive-definite kernel function $k(u, v)$ satisfies the Cauchy-Schwartz inequality [4 pts]

In the case of 2×2, the Gram matrix is

$$K = \begin{bmatrix} k(u,u) & k(u,v) \\ k(v,u) & k(v,v) \end{bmatrix}$$

and the Cauchy-Schwartz inequality is

$$k(u,v)^2 \leq k(u,u)k(v,v)$$

(c) Gaussian Kernel [10 pts]

Given the Gaussian Kernel

$$k(u,v) = \exp(-\|u-v\|^2/2\sigma^2)$$

By making use of the expansion

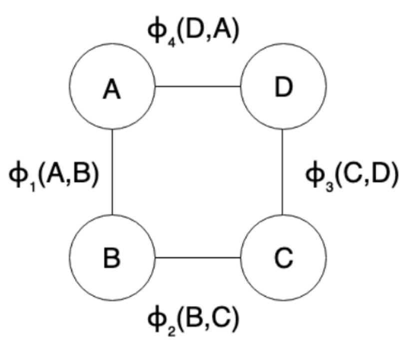
$$k(u,v) = \exp(-u^{\mathrm{T}}u/2\sigma^2)\exp(u^{\mathrm{T}}v/\sigma^2)\exp(-v^{\mathrm{T}}v/2\sigma^2)$$

and then expanding the middle factor as a power series, show that the Gaussian kernel can be expressed as the inner product of an infinite-dimensional feature vector.

2 Markov Random Field [20 pts]

TA: Pranathi B. Suresha

(a) Given an undirected graph G over variables X = A,B,C,D as given in the image below, calculate the joint probability Pr(A,B,C,D) as normalized product of factors. Show both the unnormalized and normalized product of factors in your solution[10 pts]



$$Hint : Pr(A = a, B = b, C = c, D = d) = \frac{\phi_1(a,b)\phi_2(b,c)\phi_3(c,d)\phi_4(d,a)}{\sum_{a'}\sum_{b'}\sum_{c'}\sum_{d'}\phi_1(a',b')\phi_2(b',c')\phi_3(c',d')\phi_4(d',a')} \tag{1}$$

The factors $\phi_1(A,B)$, $\phi_2(B,C)$, $\phi_3(C,D)$ and $\phi_4(D,A)$ are as given below:

A	B	$\phi_1(A,B)$
a^0	b^0	5
a^0	b^1	10
a^1	b^0	15
a^1	b^1	20

B	C	$\phi_2(B,C)$
a^0	b^0	30
a^0	b^1	40
a^1	b^0	50
a^1	b^1	60

A	B	$\phi_3(C,D)$
a^0	b^0	25
a^0	b^1	50
a^1	b^0	75
a^1	b^1	100

B	C	$\phi_4(D,A)$
a^0	b^0	150
a^0	b^1	200
a^1	b^0	250
a^1	b^1	300

(b) Answer the following questions. Please give only one line explanation. [10 pts]

- What is a clique and what is a maximal clique? [2 pts]
- Write one difference between Directed graphical model (DGM)/Bayes Networks (BN) and Undirected Graphical Models (UGM)/Markov Networks(MN)? [2 pts]
- In Markov random fields what is the reason for using a potential function instead of a probability function? [2 pts]
- Write the exponential form of Probability distribution $P(X_1, ..., X_n)$ for pairwise markov networks [2 pts]
- Write one application of pairwise Markov random fields in computer vision.

3 Hidden Markov Model [10 pts]

TA: Chukang Zhong

This problem will let you get familiar with HMM algorithms by doing the calculations by hand. There are three coins $(1, 2, 3)$, to throw them randomly, and record the result. $S = 1, 2, 3$; $V = H, T$ (Head or Tail); A, B, π is given as

A:

	1	2	3
1	0.9	0.05	0.05
2	0.45	0.1	0.45
3	0.45	0.45	0.1

π :

π	1/3	1/3	1/3

B:

	1	2	3
H	0.5	0.75	0.25
T	0.5	0.25	0.75

(a) Given the model above, what's the probability of observation $O = H, T, H$. [10 pts]

4 Neural networks [20 pts]

TA: Chukang Zhong

(a) Consider a neural network for a binary classification using sigmoid function for each unit. If the network has no hidden layer, explain why the model is equivalent to logistic regression. [5 pts]

(b) Consider a simple two-layer network. Given the cost function used to train the neural network

$$\ell(w, \alpha, \beta) = \sum_{i=1}^{m} \left(y^i - \sigma\left(w^T z^i\right)\right)^2$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. Show that the gradient is given by

$$\frac{\partial \ell(w, \alpha, \beta)}{\partial w} = \sum_{i=1}^{m} 2\left(y^i - \sigma\left(u^i\right)\right) \sigma\left(u^i\right)\left(1 - \sigma\left(u^i\right)\right) z^i$$

where $z_1^i = \sigma\left(\alpha^T x^i\right)$, $z_2^i = \sigma\left(\beta^T x^i\right)$. Also find the gradient of ℓ with respect to α and β. [15 pts]

5 Programming [30 pts]

TA: Zongen Li

In this problem, you will implement algorithm to analyze the behavior of *SP500* index over a period of time. For each week, we measure the price movement relative to the previous week and denote it using a binary variable (+1 indicates up and -1 indicates down). The price movements from week 1 (the week of January 5) to week 39 (the week of September 28) are included in `sp500.mat`.

Consider a Hidden Markov Model in which x_t denotes the economic state (good or bad) of week t and y_t denotes the price movement (up or down) of the *SP500* index. We assume that $x_{(t+1)} = x_t$ with probability 0.8, and $P_{(Y_t|X_t)}(y_t = +1|x_t = \text{good}) = P_{(Y_t|X_t)}(y_t = -1|x_t = \text{bad}) = q$. In addition, assume that $P_{(X_1)}(x_1 = \text{bad}) = 0.8$. Load the `sp500.mat`, implement the algorithm in `algorithm.m/algorithm.py` and submit this file. In your report, briefly describe how you implement your algorithm and report the following :

(a) Assuming $q = 0.7$, plot $P_{(X_t|Y)}(x_t = \textbf{good}|y)$ for $t = 1, 2, ..., 39$. What is the probability that the economy is in a good state in the week of week 39. [15 pts]

(b) Repeat (a) for $q = 0.9$, and compare the result to that of (a). Explain your comparison in one or two sentences. [15 pts]

6 Extra credits : Support Vector Machines [20 pts]

TA: Pranathi B Suresha

Recall that in class, we talked about soft margin SVM:

$$\begin{aligned} \min_{w,b,\xi} \quad & ||w||^2 + C\sum_{i}^{m} \xi^i \\ s.t. \quad & y^i(w^\top x^i + b) \geq 1 - \xi^i, \quad \xi^i \geq 0, \quad \forall i \end{aligned} \tag{2}$$

Let $(x^i, y^i)_{i=1}^m$ with $x^i \in R^n$ and $y^i \in (\pm 1), i \in [1 : m]$, be a linearly separable set of training data. Show that if C is sufficiently large, the solution of the primal soft SVM problem will give the unique maximum margin separating hyperplane. How large does C need to be?
(Hint: (1) The hard margin SVM has a unique solution for the linearly separable set, but the soft margin SVM (eq. 2) doesn't give a unique solution for an arbitrary C unless C satisfies some condition, which is what the question asks. For more information, please refer to this paper. (2) In order to obtain the same unique hyperplane as the solution to the hard SVM, we need to make all slack variables $\xi^i = 0$. In that case, the soft SVM becomes the hard SVM. (3) derive KKT conditions for soft margin SVM and use KKT conditions to find C so that hint (2) is satisfied).