CSE 6740B HW4

Che-Ting,Meng

November 17th,2022

1 Kernels

1.1 (a)

Kernel functions have to satisfy being symmetric and are positive semi-definite. The former is satisfied in the given case and the latter requires $\forall v R^m, v^T K v \geqslant 0$.

1.1.1 1

False / Not definite.
$v^T K v = v^T (k_1 - k_2) v = v^T k_1 v - v^T k_2 v$
$\because v^T k_1 v > 0, v^T k_2 v > 0$ (positive definite),
$\therefore v^T K v$ can be any arbitrary real number matrix, it is uncertain whether it will be positive semidefinite and subsequently not necessarily a kernel function.

1.1.2 2

True.
$v^T K v = v^T (k_1 * k_2) v = v^T k_1 v * v^T k_2 v$
$\because v^T k_1 v > 0, v^T k_2 v > 0$ (positive definite),
$\therefore v^T K v > 0$, it will be positive definite and subsequently a kernel function.

1.1.3 3

True.
For all real numbers in the kernel matrix, $exp(\gamma||u - v||^2) \geqslant 0$, and K(u,v) can be written in LDU form $K = L * D * U$. Also, because K is symmetrical, it satisfies $L = U^T$:

$$K = LDL^T = LD * DL^T = (LD)(LD)^T$$
$$v^T K v = (vLD)^T(vLD) = y^T y = ||y||^2 \geqslant 0$$

1.2 (b)

$\because$ Positive definite matrices have positive determinants,
$\therefore det(K) = k(u,u)k(v,v) - k(u,v)^2 > 0$
proving: $k(u,v)^2 \leqslant k(u,u)k(v,v)$

1.3 (c)

Expanding the middle factor: $exp(u^T v/\sigma^2) = \Sigma_{n=0}^{\infty} \frac{(u^T v)^n}{n!}$,then plug it back.
Denoting $exp(u^T u/2\sigma^2)$ as c_u and $exp(v^T v/2\sigma^2)$ as c_v, which are scalars:
k(u,v)=$\Sigma_{n=0}^{\infty} \frac{[(c_u * u^T) * (c_v * v)]^n}{n!}$, which is the inner product of an indefinite-dimensional feature vector, or a linear combination of polynomial of infinite possible features.

2 Markov Random Field

2.1 (a)

	Unnormalized	Normalized
a0 b0 c0 d0	562500	0.003761
a0 b0 c0 d1	1875000	0.012537
a0 b0 c1 d0	2250000	0.015044
a0 b0 c1 d1	5000000	0.033431
a0 b1 c0 d0	1875000	0.012537
a0 b1 c0 d1	6250000	0.041789
a0 b1 c1 d0	6750000	0.045132
a0 b1 c1 d1	15000000	0.100293
a1 b0 c0 d0	2250000	0.015044
a1 b0 c0 d1	6750000	0.045132
a1 b0 c1 d0	9000000	0.060176
a1 b0 c1 d1	18000000	0.120351
a1 b1 c0 d0	5000000	0.033431
a1 b1 c0 d1	15000000	0.100293
a1 b1 c1 d0	18000000	0.120351
a1 b1 c1 d1	36000000	0.240702

2.2 (b)

2.2.1

Clique
A subset of vertices such that every two of them are adjacent, implying that the subgraph is complete.
Maximal clique
A clique that cannot be extended by including one more adjacent vertex, meaning it is not a subset of a larger clique.

The fact that cliques are complete and that maximal cliques cannot be further extended allow us to write each maximal clique's potential as the product of all its sub-cliques because each clique is complete and independent.

2.2.2

DGM/BN VS. UGM/MN:
DGM/BN assumes causalities in nodes where UGM/MN only assumes relationships or information flows. So in one way, DGM/BN is of stronger assumption. This leads to directions in DGM/BN and causes different Markov blanket in DGM/BN and UGM/MN. The parent nodes are not independent of each other given the mutual children node in DGM/BN, where as for UGM/MN, if the mutual shared node is given, the connected nodes are independent of each other.

2.2.3

It is because in Markov Random Fields, there are no parent nodes. So they don't have to use probability function to represent conditional distribution as in directed markov models.

2.2.4

$P(X_1, ..., X_n) = \frac{1}{Z} exp(\Sigma_{i,j \in E} \theta_{ij} X_i X_j + \Sigma_{i \in V} \theta_i X_i + \Sigma_{i \in V} \alpha_i X_i^2)$

2.2.5

It is used in synthetic noisy image problems.

3 Hidden Markov Model

3.1 (a)

For joint distribution factorization:
$P(X_1 = H, X_2 = T, X_3 = H, Y_1, Y_1, Y_3) = P(Y_1) \sqcap_{i,j} A_{ij}^{y_t^i y_{t+1}^j} \sqcap_i \sqcap_k B_{ik}^{y_t^i x_i^k}$
Calculate summation of the possibilities of observation H,T,H from 27 potential state combinations.

$$
\begin{aligned}
P((X_1 = H, X_2 = T, X_3 = H) &= \Sigma_Y P(X_1 = H, X_2 = T, X_3 = H, Y_1, Y_1, Y_3) \\
&= \pi \Sigma_Y \sqcap_{i,j} A_{ij}^{y_t^i y_{t+1}^j} \sqcap_i \sqcap_k B_{ik}^{y_t^i x_i^k} \\
&= \frac{1}{3}[0.9^2 * 0.5^3 + 0.9 * 0.05 * 0.5^2 * 0.75 + ... + 0.1^2 * 0.25^2 * 0.75] \\
&= 0.13922 = 13.922\%
\end{aligned}
$$

4 Neural Network

4.1 (a)

Without any hidden layer, the neural network would take input layer x_is, assign weights w_{ij}s to them and pass the linear combinations $\Sigma_i w_{ij} x_i$ to the output layer. Then the summation would go through the Sigmoid function: f(u)=$\frac{1}{1+exp(-u)}$, which in this case, u=$\Sigma_i w_{ij} x_i$. Finally, the neural network uses gradient descend to find the best parameter.
And the above is equivalent to the process of a linear logistic regression.

4.2 (b)

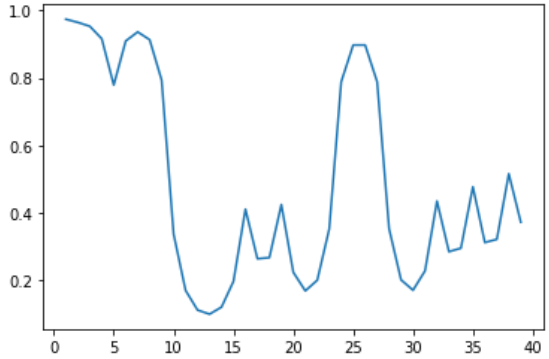
$$
\begin{aligned}
l(w, \alpha, \beta) &= \Sigma_i^m (y^i - \sigma(w^T z^i))^2 \\
&= \Sigma_i^m (y^i - \sigma(w^T \sigma(\beta^T \sigma(\alpha^T x^i))))^2 \\
\frac{\partial l}{\partial w} &= \Sigma_i^m 2(y^i - \sigma(w^T z^i)) * (\frac{\partial \sigma}{\partial w}) z^i \\
&= \Sigma_i^m 2[y^i - \sigma(w^T z^i)] * [\sigma(w^T z^i)(1 - \sigma(w^T z^i)] z^i \\
&\quad where(u^i = w^T z^i) \\
&= \Sigma_i^m 2(y^i - \sigma(u^i))\sigma(u^i)(1 - \sigma(u^i)) z^i \\
\frac{\partial l}{\partial \beta} &= \frac{\partial l}{\partial u^i} \frac{\partial u^i}{\partial z_2^i} \frac{\partial z_2^i}{\partial \beta} \\
&= \Sigma_i^m 2(y^i - \sigma(u^i))\sigma(u^i)(1 - \sigma(u^i) * w^T * \sigma(z_2^i)(1 - \sigma(z_2^i)) z_1^i \\
\frac{\partial l}{\partial \alpha} &= \frac{\partial l}{\partial u^i} \frac{\partial u^i}{\partial z_2^i} \frac{\partial z_2^i}{\partial z_1^i} \frac{\partial z_1^i}{\partial \alpha} \\
&= \Sigma_i^m 2(y^i - \sigma(u^i))\sigma(u^i)(1 - \sigma(u^i) * w^T * \sigma(z_2^i)(1 - \sigma(z_2^i)) * \beta^T * \sigma(z_1^i)(1 - \sigma(z_1^i)) x^i
\end{aligned}
$$

5 Programming

This problems is looking for $P(x_3 9 = good|y_3 9)$. According to the bayes rules, we can calculate it with $\frac{P(x_{39}^{good}=1,y_{39})}{P(y_{39})}$, which can be computed using a combination of forward and backward algorithm $\frac{\alpha_{39}^{good}\beta_{39}^{good}}{P(y_{39})}$. I created two functions–forward and backward, in the algo function to simulate the HMM operation.

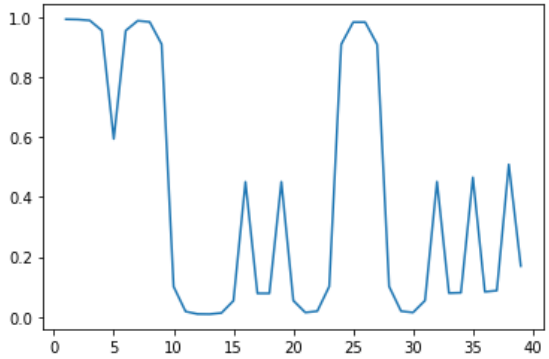
5.1 (a)

The probability for a good economy state in week 39 with q=0.7 is 62.73%.



5.2 (b)

The probability for a good economy state in week 39 with q=0.9 is 83.02%.



Discussion: It is clear that the results with different emission probabilities show essentially similar patterns, only with different magnitude. Parameter q=0.9 has an amplified magnitude. This is because the same transition probabilities and starting

probability, the hidden state structure stays the same, which would maintain the graphical pattern. Meanwhile, a larger emission probability enables the model to reach more extreme probabilities at the same time points.

6 Extra credits : Support Vector Machines

Derive Lagrangian Multiplier:

$$\mathrm{L}(w, w_0, \zeta, \alpha, \mu)=\frac{1}{2}||w||^2 + C\Sigma_i^N\zeta_i - \Sigma_i^N\alpha_i(y_i(w^T\phi(x_i) + w_0) - 1 + \zeta_i) - \Sigma_i^N\mu_i\zeta_i)$$

Define KKT conditions:
1.$\alpha_i \geqslant 0$
2.$y_i(w^T\phi(x_i) + w_0) - 1 + \zeta_i \geqslant 0$
3.$\alpha_i(y_i(w^T\phi(x_i) + w_0) - 1 + \zeta_i) = 0$
4.$\mu_i \geqslant 0$
5.$\zeta_i \geqslant 0$
6.$\mu_i\zeta_i = 0$

From derivation $\frac{\partial L}{\partial \zeta_i} = 0$, we can get $\alpha_i = C - \mu_i$.
$\because \alpha_i \geqslant 0$, $\mu_i \geqslant 0$,
$\therefore C \geqslant \alpha_i \geqslant 0$;
For support vectors, $\alpha_i \neq 0$ and $y_i(w^T\phi(x_i) + w_0) = 1 - \zeta_i$. If we transform the SVM into a hard margin SVM and make $\zeta_i = 0$, then:
$\alpha_i > 0$ and to ensure $y_i(w^T\phi(x_i) + w_0) = 1 \rightarrow \mu_i \neq 0 \rightarrow \mu_i > 0$,
$\therefore$ C has to satisfy $C > \alpha_i > 0$ to guarantee a hyperplane unique solution.