CSE 6740 - HW 4

Manvitha Kalicheti
gtID: 903838438

December 8, 2022

Collaborators

Keerthan Ramnath, Aishwarya Vijaykumar Sheelvant

References

Pattern Recognition and Machine Learning by Christopher Bishop, Class slides

1 Kernels

(a)

1.

False.

$k(u, v) = k_1(u, v) - k_2(u, v)$

To check if k is a valid kernel, we have to check if its Gram matrix, K is PSD.

$K = K_1 - K_2$

We know that K_1 and K_2 are PSD, i.e., $zK_1z^T > 0$ and $zK_1z^T > 0$ $\forall z \in \mathcal{R}^m$.

$zKz^T = zK_1z^T - zK_2z^T$

We cannot say that $zK_1z^T > zK_2z^T$ $\forall z \in \mathcal{R}^m$.

Thus, we cannot say that $zKz^T > 0$.

$\implies$ k might not be a valid kernel function.

2.

True.

Since k_1 and k_2 are valid kernels, let ϕ_1 be the feature map for k_1 and ϕ_2 be the feature map for k_2. Then, $k_1(u, v) = \phi_1(u).\phi_1(v)$ and $k_2(u, v) = \phi_2(u).\phi_2(v)$. Let $f_i(u)$ and $g_i(u)$ be the ith feature value under feature map ϕ_i and ϕ_2 respectively.

$$k(u, v) = k_1(u, v)k_2(u, v)$$

$$
\begin{aligned}
&= (\phi_1(u).\phi_1(v))(\phi_2(u).\phi_2(v)) \\
&= \left(\sum_i f_i(u)f_i(v)\right)\left(\sum_j g_j(u)g_j(v)\right) \\
&= \sum_{i,j} f_i(u)f_i(v)g_j(u)g_j(v) \\
&= \sum_{i,j}(f_i(u)g_j(u))(f_i(v)g_j(v)) \\
&= \sum_{i,j}(h_{i,j}(u))(h_{i,j}(v)) \\
&= \phi_3(u).\phi_3(v)
\end{aligned}
$$

where we define a feature map ϕ_3 with a feature value $h_{i,j}(u)$ defined as $h_{i,j}(u) = f_i(u)g_j(u)$.

We now have $k(u,v) = \phi_3(u).\phi_3(v)$ where the inner product sums over all pairs $< i,j >$. Thus $k(u,v)$ is a valid kernel.

3.

False.

Consider the gram matrix for 2 points u and v: $K = \begin{bmatrix} 1 & k(u,v) \\ k(v,u) & 1 \end{bmatrix}$

If K is PSD, the product of its eigen values > 0, i.e., $|K| > 0$. As $k(u,v) = k(v,u)$ from symmetry, we get

$$
\begin{aligned}
1 - k^2(u,v) &> 0 \\
\implies k^2(u,v) &< 1 \\
\implies \exp(2\gamma||u-v||^2) &< 1 \\
\implies 2\gamma||u-v||^2 &< 0 \text{ (applying log on both sides)}
\end{aligned}
$$

But $\gamma > 0$. This means that the above expression cannot be negative. $\implies k(u,v) = \exp(\gamma||u-v||^2)$ with $\gamma > 0$ is not a valid kernel.

(b)

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

Since K is a gram matrix made of a valid kernel k, K is PSD which means its eigen values are > 0 and $k(u,v) = k(v,u)$.

$$
\begin{aligned}
|K - \lambda I| &= 0 \\
(k(u,u) - \lambda)(k(v,v) - \lambda) - k(u,v)k(v,u) &= 0 \\
\lambda^2 - \lambda(k(u,u) + k(v,v)) + k(u,u)k(v,v) - k^2(u,v) &= 0
\end{aligned}
$$

As we know that λ_1 and λ_2 are both > 0, $\lambda_1\lambda_2$ ¿ 0 $\implies$ $k(u,u)k(v,v) < k^2(u,v)$. This is the Cauchy-Schwartz inequality.

(c)

$$\begin{aligned} k(u,v) &= \exp\left(-\frac{||u-v||^2}{2\sigma^2}\right) \\ &= \exp\left(-\frac{u^Tu + v^Tv - 2u^Tv}{2\sigma^2}\right) \\ &= \exp\left(-\frac{u^Tu}{2\sigma^2}\right)\exp\left(-\frac{v^Tv}{2\sigma^2}\right)\exp\left(\frac{u^Tv}{\sigma^2}\right) \end{aligned}$$

Applying Taylor series expansion on the last term:

$$\begin{aligned} \exp\left(\frac{u^Tv}{\sigma^2}\right) &= 1 + \frac{u^Tv}{1!\sigma^2} + \frac{(u^Tv)^2}{2!\sigma^4} + \frac{(u^Tv)^3}{3!\sigma^6} + \dots \\ &= \frac{\sqrt{u^{T0}}}{\sigma^0}\frac{\sqrt{v^0}}{\sigma^0} + \frac{\sqrt{u^{T1}}}{\sqrt{1!}\sigma^1}\frac{\sqrt{v^1}}{\sqrt{1!}\sigma^1} + \frac{\sqrt{(u^T)^2}}{\sqrt{2!}\sigma^2}\frac{\sqrt{(v)^2}}{\sqrt{2!}\sigma^2}\dots \end{aligned}$$

Substituting this back in the previous equation

$$k(u,v) = \exp\left(-\frac{u^Tu}{2\sigma^2}\right)\exp\left(-\frac{v^Tv}{2\sigma^2}\right)\left(\frac{\sqrt{u^{T0}}}{\sigma^0}\frac{\sqrt{v^0}}{\sigma^0} + \frac{\sqrt{u^{T1}}}{\sqrt{1!}\sigma^1}\frac{\sqrt{v^1}}{\sqrt{1!}\sigma^1} + \frac{\sqrt{(u^T)^2}}{\sqrt{2!}\sigma^2}\frac{\sqrt{(v)^2}}{\sqrt{2!}\sigma^2}\dots\right)$$

Let's define an infinite dimensional feature vector

$$\phi(x) = \exp\left(-\frac{x^Tx}{2\sigma^2}\right)\left(\frac{\sqrt{x^{T0}}}{\sigma^0}, \frac{\sqrt{x^{T1}}}{\sqrt{1!}\sigma^1}, \frac{\sqrt{(x^T)^2}}{\sqrt{2!}\sigma^2}\dots\right)$$

Then,

$$k(u,v) = \phi(u)^T\phi(v)$$

Hence, Gaussian kernel is an inner product of infinite dimensinal feature vector.

2 Markov Random Field

(a)

Assignment				Unnormalised	Normalised
a	b	c	d		
0	0	0	0	56250	0.0038
0	0	0	1	1875000	0.0125
0	0	1	0	2250000	0.0150
0	0	1	1	5000000	0.0334
0	1	0	0	1875000	0.0125
0	1	0	1	6250000	0.0418
0	1	1	0	6750000	0.0451
0	1	1	1	15000000	0.1003
1	0	0	0	2250000	0.0150
1	0	0	1	6750000	0.0451
1	0	1	0	9000000	0.0602
1	0	1	1	18000000	0.1204
1	1	0	0	5000000	0.0334
1	1	0	1	15000000	0.1003
1	1	1	0	18000000	0.1204
1	1	1	1	36000000	0.2407
				149562500	1.0000

Table 1: Joint Probability Pr(A,B,C,D) as normalized product of factors

(b)

(1.)

A clique, C, in an undirected graph $G = (V, E)$ is a subset of the vertices, $C \subseteq V$, such that every two distinct vertices are adjacent, implying that the induced subgraph is complete.

A maximal clique is a clique that cannot be extended by including one more adjacent vertex, meaning it is not a subset of a larger clique.

(2.)

The directed edges in a DGM/BN give causality relationships, the undirected edges in a UGM/MN give correlations between variables.

(3.)

The reason for using a potential function instead of a probability function is becausee in MRFs there does not exist a parent.

(4.)

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

(5.)

Pairwise Markov random fields are used for image segmentation in computer vision.

3 Hidden Markov Models

First, we need H:

$$\begin{aligned}
P(H, S1) &= \pi_1 b_{H,1} \\
&= \frac{1}{3} 0.5 = 0.1667 \\
P(H, S2) &= \pi_2 b_{H,2} \\
&= \frac{1}{3} 0.75 = 0.25 \\
P(H, S3) &= \pi_3 b_{H,3} \\
&= \frac{1}{3} 0.25 = 0.0833
\end{aligned}$$

Next, we need T:

$$\begin{aligned}
P(HT, S1) &= [P(H, S1)a_{1,1} + P(H, S2)a_{2,1} + P(H, S3)a_{3,1}]b_{T,1} \\
&= [0.9 \times 0.1667 + 0.45 \times 0.25 + 0.45 \times 0.0833] \times 0.5 \\
&= 0.15 \\
P(HT, S2) &= [P(H, S1)a_{1,3} + P(H, S2)a_{2,2} + P(H, S3)a_{3,2}]b_{T,2} \\
&= [0.05 \times 0.1667 + 0.1 \times 0.25 + 0.45 \times 0.0833] \times 0.25 \\
&= 0.0177 \\
P(HT, S3) &= [P(H, S1)a_{1,3} + P(H, S2)a_{2,3} + P(H, S3)a_{3,3}]b_{T,3} \\
&= [0.05 \times 0.1667 + 0.45 \times 0.25 + 0.1 \times 0.0833] \times 0.75 \\
&= 0.0969
\end{aligned}$$

Next, we need H again:

$$\begin{aligned}
P(HTH, S1) &= [P(HT, S1)a_{1,1} + P(HT, S2)a_{2,1} + P(HT, S3)a_{3,1}]b_{T,1} \\
&= [0.15 \times 0.9 + 0.01771 \times 0.45 + 0.09687 \times 0.45] \times 0.5 \\
&= 0.0932 \\
P(HTH, S2) &= [P(HT, S1)a_{1,3} + P(HT, S2)a_{2,2} + P(HT, S3)a_{3,2}]b_{T,2} \\
&= [0.15 \times 0.05 + 0.01771 \times 0.01 + 0.09687 \times 0.45] \times 0.75 \\
&= 0.0396 \\
P(HTH, S3) &= [P(HT, S1)a_{1,3} + P(HT, S2)a_{2,3} + P(HT, S3)a_{3,3}]b_{T,3} \\
&= [0.15 \times 0.05 + 0.01771 \times 0.45 + 0.09687 \times 0.1] \times 0.25 \\
&= 0.0063
\end{aligned}$$

Finally, the probability of the observation:

$$\begin{aligned}
P(HTH) &= \sum_i P(HTH, Si) \\
&= \boxed{0.1391}
\end{aligned}$$

4 Neural Networks

(a)

If the network has no hidden layer, then the loss function becomes $l(w) = \sum_{i=1}^{n}(y^i - \sigma(w^T x^i))^2$. This is the loss function of logistic regression. Thus, the model with no hidden layer becomes equivalent to logistic regression.

(b)

For the sigmoid function, $\frac{\partial \sigma(x)}{\partial x} = \sigma(x)(1 - \sigma(x))$.

$$\begin{aligned}\frac{\partial l(w, \alpha, \beta)}{\partial w} &= \sum_{i=1}^{m} 2(y^i - \sigma(w^T z^i))(-\sigma(w^T z^i)(1 - (\sigma(w^T z^i)))z^i \\ &= -\sum_{i=1}^{m} 2(y^i - \sigma(u^i))\sigma(u^i)(1 - \sigma(u^i))z^i\end{aligned}$$

Define $w^T = (w_1, w_2)$

Next, derivative wrt α:

$$\begin{aligned}\frac{\partial l(w, \alpha, \beta)}{\partial \alpha} &= \frac{\partial l}{\partial z_1^i}\frac{z_1^i}{\partial \alpha} \\ &= -\sum_{i=1}^{m} 2(y^i - \sigma(u^i))\sigma(u^i)(1 - \sigma(u^i))w_1\sigma(\alpha^T x^i)(1 - \sigma(\alpha^T x^i))x^i\end{aligned}$$

Next, derivative wrt β:

$$\begin{aligned}\frac{\partial l(w, \alpha, \beta)}{\partial \beta} &= \frac{\partial l}{\partial z_2^i}\frac{z_2^i}{\partial \beta} \\ &= -\sum_{i=1}^{m} 2(y^i - \sigma(u^i))\sigma(u^i)(1 - \sigma(u^i))w_2\sigma(\beta^T x^i)(1 - \sigma(\beta^T x^i))x^i\end{aligned}$$

5 Programming

(a)

The probability that the economy is in a good state in week 39 with $q = 0.7$ is 0.6830.

(b)

The probability that the economy is in a good state in week 39 with $q = 0.9$ is 0.8379. q denotes the probability of price rise in a good economic state week or the price drop in a bad economic state week. Thus, as expected, increasing q value further increases the peaks, and further decreases the troughs of the $P_{(X_t|Y)}(x_t = \text{good}|y)$ values as evident from the plots.

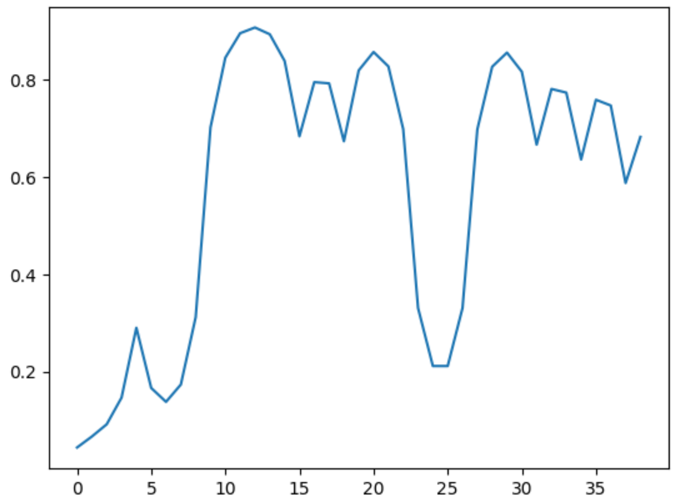

Figure 1: q = 0.7, $P_{(X_t|Y)}(x_t = \text{good}|y)$ vs $t = 1, 2, \ldots, 39$ weeks

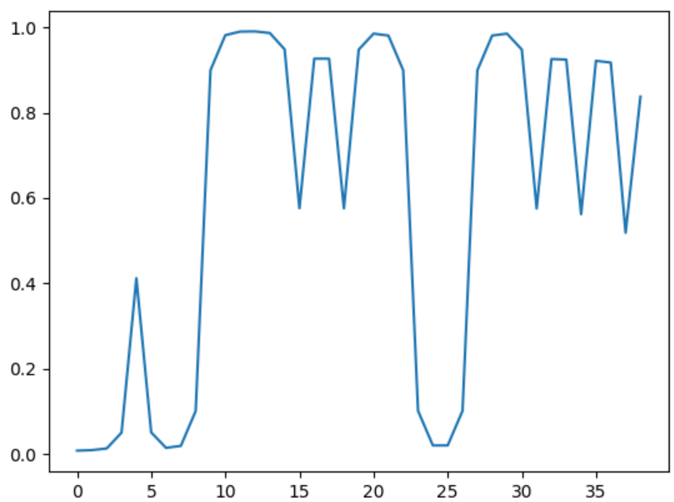

Figure 2: q = 0.9, $P_{(X_t|Y)}(x_t = \text{good}|y)$ vs $t = 1, 2, \ldots, 39$ weeks

6 Extra Credits: Support Vector Machines

Soft margin SVM:

$$min_{w,b,\epsilon} ||w||^2 + C \sum_{i}^{m} \epsilon^i$$

$$\text{s.t } y^i(w^Tx^i + b) \geq 1 - \epsilon^i, \epsilon^i \geq 0, \forall i$$

In the standard form:

$$min_{w,b,\epsilon} \frac{1}{2}w^Tw + C\sum_i^m \epsilon^i$$
$$\text{s.t } 1 - y^i(w^Tx^i + b) - \epsilon^i \leq 0, \epsilon \geq 0, \forall i$$

Lagrangian for this standard form:

$$L(w, \alpha, \beta) = \frac{1}{2}w^Tw + \sum_i^m C\epsilon^i + \alpha_i(1 - y_i(w^Tx^i + b) - \epsilon^i) - \beta\epsilon^i$$

Taking derivative of L wrt w, b and ϵ^i and setting them to zero:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^m \alpha_i y^i x^i = 0 \implies w = \sum_{i=1}^m \alpha_i y^i x^i$$
$$\frac{\partial L}{\partial b} = \sum_{i=1}^m \alpha_i y^i = 0$$
$$\frac{\partial L}{\partial \epsilon^i} = C - \alpha_i - \beta_i = 0 \implies C = \alpha_i + \beta_i$$

Putting these values back in L:

$$L(w, \alpha, \beta) = \frac{1}{2}\left(\sum_{i=1}^m \alpha_i y^i x^i\right)^T \left(\sum_{i=1}^m \alpha_i y^i x^i\right) + \sum_{i=1}^m \alpha_i(1 - y^i((\sum_{j=1}^m \alpha_j y^j x^j)^T x^i + b))$$

Simplifying:

$$L(w, \alpha, \beta) = \sum_i^m \alpha_i - \frac{1}{2}\sum_{i,j}^m \alpha_i\alpha_j y^i y^j (x^{iT}x^j)$$

KKT conditions for L:

$$y^i(w^Tx^i + b) - 1 + \epsilon^i \geq 0$$
$$\alpha_i[y^i(w^Tx^i + b) - 1 + \epsilon^i] = 0$$
$$\alpha_i \geq 0$$
$$\beta_i \geq 0$$
$$\epsilon_i \geq 0$$
$$\beta_i\epsilon^i = 0$$

We also know:

$$\alpha_i + \beta_i = C$$
$$\beta_i \geq 0$$
$$\implies \alpha_i \leq C$$

If $\alpha_i < C$, $\beta_i > 0 \implies \epsilon^i = 0$ (complementary slackness. Our problem now reduces to hard-margin SVM and we have a unique solution.

Thus, to have unique solution, soft margin SVM should satisfy

$$C > \alpha_i$$
$$\implies C > max(\alpha_1, \alpha_2, \ldots, \alpha_m)$$