FINAL PRACTICE SOLUTIONS # 1

Friday, 8th June 2018

Problem 1 (Hidden Markov Models)

(a) 0.3*0.9*0.6*0.5 =0.081

(b) The most likely assignment given $O_1 = A, O_2 = B$ is $X_1 = 0, X_2 = 1$ You can use Viterbi, or list all four possibilities.

(c) False. Conditional independent does not imply independent.

Problem 2 (EM algorithm and Gaussian Mixture Model)

(a) Using the Bayes' Rule, we get that

$$P(\text{component } 1|x) = \frac{(1/2)\mathcal{N}(x|\mu, 1)}{(1/2)\mathcal{N}(x|\mu, 1) + (1/2)\mathcal{N}(x|\mu, 2^2)}$$

Applying the three observations, we get

$$r_{11} = \frac{(1/2)0.40}{(1/2)0.40 + (1/2)(1/2)0.40} = 2/3$$
$$r_{11} = \frac{(1/2)0.33}{(1/2)0.33 + (1/2)(1/2)0.38} = 33/52$$
$$r_{31} = \frac{(1/2)0.05}{(1/2)0.05 + (1/2)(1/2)0.24} = 5/17$$

(b) First, note that the gradient of normal density function with mean μ and variance σ^2 with respect to mean μ is:

$$\frac{\partial \mathcal{N}(x_i|\mu, \sigma^2)}{\partial \mu} = \mathcal{N}(x_i|\mu, \sigma^2) \cdot \frac{x_i - \mu}{\sigma^2}$$

The log likelihood is

$$\log p(X|\mu) = \sum_{i=1}^{3} \log \left[\frac{1}{2}\mathcal{N}(x_i|\mu, 1) + \frac{1}{2}\mathcal{N}(x_i|\mu, 2^2) \right]$$

Take the derivative with respect to μ and set to zero:

$$\begin{aligned}
\frac{\partial \log p(X|\mu)}{\partial \mu} &= \sum_{i=1}^{3} \frac{1}{\frac{1}{2}\mathcal{N}(x_i|\mu,1)+\frac{1}{2}\mathcal{N}(x_i|\mu,2^2)} \left[\mathcal{N}(x_i|\mu,1)\cdot\frac{x_i-\mu}{1^2}+\mathcal{N}(x_i|\mu,2)\cdot\frac{x_i-\mu}{2^2}\right] \\
&= \sum_{i=1}^{3}\left[r_{i1}(x_i-\mu)+r_{i2}\cdot\frac{x_i-\mu}{4}\right] \\
&= \sum_{i=1}^{3}\left[\left(r_{i1}+\frac{1-r_{i1}}{4}\right)x_i-\left(r_{i1}+\frac{1-r_{i1}}{4}\right)\mu\right]=0 \\
\hat{\mu} &= \frac{\sum_{i=1}^{3}(r_{i1}+(1-r_{i1})/4)x_i}{\sum_{i=1}^{3}(r_{i1}+(1-r_{i1})/4)} = \frac{(3/4)4.0+(151/208)4.6+(25/68)2.0}{(3/4)+(151/208)+(25/68)}
\end{aligned}$$

Problem 3 (PCA)

(a) First step is zero centering however the four points are already zero centred. The matrix

$$X = \begin{bmatrix} -1 & 1 \\ 2 & 2 \\ -2 & -2 \\ 1 & -1 \end{bmatrix}$$

The goal is to find the eigenvector of $X^T X$ corresponding to the largest eigenvalue.

$$X^T X = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

$\det(X^T X - \lambda I) = \lambda^2 - 20\lambda + 64$. The eigen values are $\lambda = 16, 4$ and the eigen vector for $\lambda = 16$ can be found by solving $(X^T X - 16I)\mathbf{p} = 0$ where $\mathbf{p} \in \mathbb{R}^2$. Solving the equations we get $p_1 = p_2$ and the eigenvector is therefore $\mathbf{p} = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^T$

Note that you could have solved this geometrically, simply by noticing from the scatter plot that the direction of maximum variance is indeed the $\mathbf{p}$ that has been obtained.

(b) The new coordinates are

$$\begin{aligned}
z_1 &= x_1 \cdot p = 0 \\
z_2 &= x_2 \cdot p = 2\sqrt{2} \\
z_3 &= x_3 \cdot p = -2\sqrt{2} \\
z_4 &= x_4 \cdot p = 0
\end{aligned}$$

The new coordinates are also zero centred, the variance is

$$\frac{\sum_i z_i^2}{4} = \frac{0+(2\sqrt{2})^2+(2\sqrt{2})^2+0}{4} = 4$$

(c) In the original space the representations are $\hat{x}_i = z_i\mathbf{p}$ so we have

$$\begin{aligned}
\hat{x}_1 &= (0,0)^T \\
\hat{x}_2 &= (2,2)^T \\
\hat{x}_3 &= (-2,-2)^T \\
\hat{x}_4 &= (0,0)^T
\end{aligned}$$

The mean reconstruction error is $\frac{1}{4}\sum_i ||x_i - \hat{x}_i||_2^2 = 1$

(d) The direction of maximum variance also called as the principal vector will now be $(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})^T$. The points x_1 and x_2 lie along this direction therefore the reconstruction error is going to be zero in this case.

Problem 4 (Kernels and SVM)

(a) i. We have $A_{ij} = K(x_1, x_2) = \Phi(x_1)^T\Phi(x_2) = \Phi(x_2)^T\Phi(x_1) = K(x_2, x_1) = A_{ji}$

ii. Let $\Phi(x_i)$ be the feature map for the i^{th} example and define the matrix $\mathbf{B} = [\Phi(x_1), \cdots, \Phi(x_n)]$. It is easy to verify that $\mathbf{A} = \mathbf{B}^T\mathbf{B}$. Then, we have $\mathbf{v}^T\mathbf{A}\mathbf{v} = (\mathbf{B}\mathbf{v})^T\mathbf{B}\mathbf{v} = ||\mathbf{B}\mathbf{v}||^2 \geq 0$

(b) i. ∞, Decrease

ii. When $C = 0$, ξ_i can be arbitrary large; therefore, the model ignores the constraints, and $w = 0$ is the optimal solution.

iii. False. When the data point is correctly classified but inside the margin, ξ_i is non-zero.

iv. True. Based on the dual representation.

v. i. $W = (1, 1)$, $b = -1$

ii. Point 1,7,8

index	x_1	x_2	y
1	0	0	$-$
2	0	-4	$-$
3	-1	-1	$-$
4	-2	-2	$-$
5	3	0	+
6	0	3	+
7	1	1	+
8	3	-1	+

Table 1: The dataset S

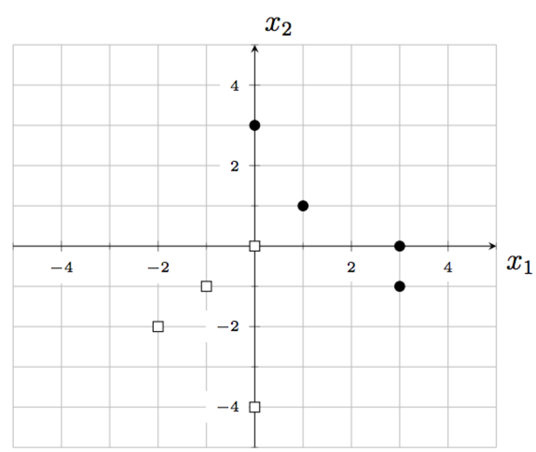

Figure 1: Linear SVM

Problem 5 (Kernelized logistic regression)

(a) By induction.
At iteration $t = 1$, this is true since $\theta = -\eta \sum_n \epsilon_n \phi(\mathbf{x}_n)$.

Assume this is true at iteration t. At iteration $t+1$, we have

$$\begin{aligned} \theta &\leftarrow \theta - \eta \sum_n \epsilon_n \phi(\mathbf{x}_n) \\ &= \sum_n \alpha_n \phi(\mathbf{x}_n) - \eta \sum_n \epsilon_n \phi(\mathbf{x}_n) \\ &= \sum_n (\alpha_n - \eta \epsilon_n) \phi(\mathbf{x}_n) \\ &= \sum_n \alpha'_n \phi(\mathbf{x}_n) \end{aligned}$$

(b)

$$\begin{aligned} h_\theta(\mathbf{x}) &= \sigma(\theta^T \phi(\mathbf{x})) \\ &= \sigma(\sum_n \alpha_n \phi(\mathbf{x}_n)^T \phi(\mathbf{x})) \end{aligned}$$

(c) If we updated $\alpha_n \leftarrow \alpha_n - \eta \epsilon_n$ and the relationship that $\theta = \sum_n \alpha_n \phi(\mathbf{x}_n)$, the corresponding update of θ would be

$$\begin{aligned} \theta &= \sum_n \alpha_n \phi(\mathbf{x}_n) \\ &= \sum_n (\alpha_n - \eta \epsilon_n) \phi(\mathbf{x}_n) \\ &= \sum_n \alpha_n \phi(\mathbf{x}_n) - \sum_n \eta \epsilon_n \phi(\mathbf{x}_n) \\ &= \theta - \eta \sum_n \epsilon_n \phi(\mathbf{x}_n) \end{aligned}$$

Problem 6 (Linear Regression)

(a)

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}) &= \mathbb{E}\left[(\boldsymbol{w}^T(\boldsymbol{x}+\boldsymbol{\epsilon}) - y)^2\right] \\ &= \mathbb{E}\left[((\boldsymbol{w}^T\boldsymbol{x} - y) + \boldsymbol{w}^T\boldsymbol{\epsilon})^2\right] \\ &= \mathbb{E}\left[((\boldsymbol{w}^T\boldsymbol{x} - y)^2 + 2(\boldsymbol{w}^T\boldsymbol{x} - y)\boldsymbol{w}^T\boldsymbol{\epsilon} + \boldsymbol{w}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\boldsymbol{w}\right] \\ &= \mathbb{E}\left[((\boldsymbol{w}^T\boldsymbol{x} - y)^2\right] + \mathbb{E}\left[2(\boldsymbol{w}^T\boldsymbol{x} - y)\boldsymbol{w}^T\right]\mathbb{E}\left[\boldsymbol{\epsilon}\right] + \boldsymbol{w}^T\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right]\boldsymbol{w} \\ &= \mathcal{L}_0(\boldsymbol{w}) + \lambda||\boldsymbol{w}||^2 \end{aligned}$$

(b) In this setting, regression assuming this type of input perturbation turns out to be equivalent to regression with an L2 regularizer.