Introduction to Machine Learning - Summer 2019
Final Exam – Solution

1 Estimation

$$
\begin{aligned}
\text{MSE}\left(\hat{\theta}\right) &\triangleq \mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]+\mathbb{E}\left[\hat{\theta}\right]-\theta\right)^2\right] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]\right)^2+2\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]\right)\left(\mathbb{E}\left[\hat{\theta}\right]-\theta\right)+\left(\mathbb{E}\left[\hat{\theta}\right]-\theta\right)^2\right] \\
&= \underbrace{\mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]\right)^2\right]}_{=V\left(\hat{\theta}\right)}+2\underbrace{\mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]\right)\right]}_{=0}\left(\mathbb{E}\left[\hat{\theta}\right]-\theta\right)+\underbrace{\left(\mathbb{E}\left[\hat{\theta}\right]-\theta\right)^2}_{b^2\left(\hat{\theta}\right)} \\
&= V\left(\hat{\theta}\right)+b^2\left(\hat{\theta}\right)
\end{aligned}
$$

2 ML

2.1

$$
\begin{aligned}
\ell(\lambda) &= \log p(\mathcal{D};\lambda) \\
&= \log\left(\prod_{i=1}^{N}\lambda x_i^{-2}\exp(-\frac{\lambda}{x_i})\right) \\
&= N\log(\lambda)-2\sum_{i=1}^{N}\log(x_i)-\lambda\sum_{i=1}^{N}\frac{1}{x_i}
\end{aligned}
$$

$$\Rightarrow \hat{\lambda}_{ML} = \arg\max_{\lambda}\ell(\lambda)$$

$$
\begin{aligned}
\frac{\partial}{\partial\lambda}\ell(\lambda) &= 0 \\
\Rightarrow \frac{N}{\lambda}-\sum_{i=1}^{N}\frac{1}{x_i} &= 0 \\
\Rightarrow \hat{\lambda}_{ML} &= \frac{N}{\sum_{i=1}^{N}\frac{1}{x_i}}
\end{aligned}
$$

2.2

$$\hat{\lambda}_{ML} = \frac{2}{2+\frac{1}{2}} = \frac{4}{5}$$

3 MAP

3.1

$$\int\limits_{-\infty}^{\infty} f_K(k;\alpha\beta)\mathrm{d}k = 1$$
$$\Rightarrow \int_0^1 C\cdot k^{\alpha-1}(1-k)^{\beta-1}\mathrm{d}k = 1$$
$$\Rightarrow C = \frac{1}{\int_0^1 k^{\alpha-1}(1-k)^{\beta-1}\mathrm{d}k}$$

3.2

$$\begin{aligned} P\left(\text{Heads and Tails}\right) &= P\left(\text{Heads}\cap\text{Tails}\right) + P\left(\text{Tails}\cap\text{Heads}\right) \\ &= 2P\left(\text{Heads}\right)P\left(\text{Tails}\right) \\ &= 2k(1-k). \end{aligned}$$

3.3

$$\begin{aligned} \hat{k}_{MAP} &= \underset{0\le k\le 1}{\arg\max}\ P\left(\text{Heads}\cap\text{Tails}\right)\cdot f_K(k;\alpha\,\beta) \\ &= \underset{0\le k\le 1}{\arg\max}\ 2k(1-k)C\cdot k^{\alpha-1}(1-k)^{\beta-1} \\ &= \underset{0\le k\le 1}{\arg\max}\ k(1-k)k^{\alpha-1}(1-k)^{\beta-1} \\ &= \underset{0\le k\le 1}{\arg\max}\ k^{\alpha}(1-k)^{\beta} \\ &= \underset{0\le k\le 1}{\arg\max}\ k^3(1-k) \end{aligned}$$

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}k}\left(k^3(1-k)\right) &= 0 \\ 3k^2(1-k)-k^3 &= 0 \\ k^2(3-3k-k) &= 0 \\ \Rightarrow \hat{k}_{MAP} &= \frac{3}{4} \end{aligned}$$

4 Non-parametric estimation

1.

$$\begin{aligned} \mathbb{E}\left[\hat{F}_X\left(x_0\right)\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{I}\left\{x_i\le x_0\right\}\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\boldsymbol{I}\left\{x_i\le x_0\right\}\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\Pr\left\{x_i\le x_0\right\} = \Pr\left\{x_1\le x_0\right\} \\ &= F_X\left(x_0\right) \end{aligned}$$

$$\Rightarrow b\left(\hat{F}_X\left(x_0\right)\right) = \mathbb{E}\left[\hat{F}_X\left(x_0\right)\right] - F_X\left(x_0\right) = 0$$

Hence, $\hat{F}_X$ is unbiased.

2. In Lecture 1, we proved that:

$$\mathrm{MSE}\left(\hat{\theta}\right) \triangleq \mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right] = b^2\left(\hat{\theta}\right) + V\left(\hat{\theta}\right)$$

$$\begin{aligned}
\Rightarrow \mathrm{MSE}\left(\hat{F}_X\left(x_0\right)\right) &= \underbrace{b^2\left(\hat{F}_X\left(x_0\right)\right)}_{=0} + \mathrm{Var}\left(\hat{F}_X\left(x_0\right)\right) = \mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^{N}\underbrace{\boldsymbol{I}\left\{x_i \le x_0\right\}}_{\triangleq Y_i}\right) \\
&= \frac{1}{N^2}\mathrm{Var}\left(\sum_{i=1}^{N} Y_i\right), \qquad \left\{Y_i\right\}_i \text{ are i.i.d} \\
&= \frac{1}{N^2} N \mathrm{Var}\left(Y_1\right) = \frac{\mathbb{E}\left[Y_1^2\right] - \mathbb{E}^2\left[Y_1\right]}{N} \\
&= \frac{F_X\left(x_0\right) - F_X^2\left(x_0\right)}{N}
\end{aligned}$$

3. Using the previous result, we have:

$$\mathrm{MSE}\left(\hat{F}_X\left(x_0\right)\right) \underset{N\to\infty}{\longrightarrow} 0$$

5 PCA I

1. By definition:

$$\boldsymbol{\mu}_y = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{U}^T\left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right) = \boldsymbol{U}^T\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right) = \boldsymbol{U}^T\left(\overbrace{\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i}^{\boldsymbol{\mu}_x} - \boldsymbol{\mu}_x\right) = 0$$

2. Consider the eigen-value decomposition: $\boldsymbol{\Sigma}_x = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ ($\boldsymbol{U}$ is unitary) Hence,

$$\begin{aligned}
\boldsymbol{\Sigma}_y &= \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{y}_i - \boldsymbol{\mu}_y\right)\left(\boldsymbol{y}_i - \boldsymbol{\mu}_y\right)^T = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y}_i\boldsymbol{y}_i^T \\
&= \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{U}^T\left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right)^T\boldsymbol{U} = \boldsymbol{U}^T\underbrace{\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right)^T\boldsymbol{U}}_{=\boldsymbol{\Sigma}_x} \\
&= \boldsymbol{U}^T\boldsymbol{\Sigma}_x\boldsymbol{U} = \boldsymbol{U}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{\Lambda}
\end{aligned}$$

3. Since $\boldsymbol{U}^T$ is unitary, we have $\left\|\boldsymbol{U}^T\boldsymbol{v}\right\|_2 = \|\boldsymbol{v}\|_2$, thus:

$$\left\|\boldsymbol{y}_i - \boldsymbol{y}_j\right\|_2 = \left\|\boldsymbol{U}^T\left(\boldsymbol{x}_i - \boldsymbol{\mu}_x\right) - \boldsymbol{U}^T\left(\boldsymbol{x}_j - \boldsymbol{\mu}_x\right)\right\|_2 = \left\|\boldsymbol{U}^T\left(\boldsymbol{x}_i - \boldsymbol{x}_j\right)\right\|_2 = \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|_2$$

6

6.1 PCA II

- (a) – (3)
- (b) – (2)
- (c) – (1)
- (d) – (4)

6.2 K-means

For fixed clusters $\{\mathcal{C}_k\}$,
we can compare the gradient with respect to $\boldsymbol{\mu}_s$ to zero:

$$\nabla_{\mu_s}\left(\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in\mathcal{C}_k}\|\boldsymbol{x}_i-\boldsymbol{\mu}_k\|_2^2\right)=\mathbf{0}$$
$$-2\sum_{\boldsymbol{x}_i\in\mathcal{C}_s}(\boldsymbol{x}_i-\boldsymbol{\mu}_s)=\mathbf{0}$$
$$\Rightarrow\boxed{\boldsymbol{\mu}_s=\frac{1}{|\mathcal{C}_s|}\sum_{\boldsymbol{x}_i\in\mathcal{C}_s}\boldsymbol{x}_i}$$

In words, the optimal centroid $\boldsymbol{\mu}_k$ of the kth cluster $\mathcal{C}_k$ is the mean of the cluster.

7 MAP classifier

$$\begin{aligned}
p(\boldsymbol{x}|C_1)\,p_{\mathcal{Y}}(C_1)&=p(\boldsymbol{x}|C_2)\,p_{\mathcal{Y}}(C_2)\\
p(\boldsymbol{x}|C_1)\,p_1&=p(\boldsymbol{x}|C_2)\,(1-p_1)\\
\frac{p_1}{1-p_1}e^{-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}_1\|_2^2}&=e^{-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}_2\|_2^2}\\
\log\left(\frac{p_1}{1-p_1}\right)-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}_1\|_2^2&=-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{\mu}_2\|_2^2\\
2\log\left(\frac{p_1}{1-p_1}\right)-\|\boldsymbol{x}\|^2+2\boldsymbol{\mu}_1^T\boldsymbol{x}-\|\boldsymbol{\mu}_1\|^2&=-\|\boldsymbol{x}\|^2+2\boldsymbol{\mu}_2^T\boldsymbol{x}-\|\boldsymbol{\mu}_2\|^2\\
2\log\left(\frac{p_1}{1-p_1}\right)+\|\boldsymbol{\mu}_2\|^2-\|\boldsymbol{\mu}_1\|^2+2\boldsymbol{\mu}_1^T\boldsymbol{x}-2\boldsymbol{\mu}_2^T\boldsymbol{x}&=0\\
\underbrace{2\log\left(\frac{p_1}{1-p_1}\right)+\|\boldsymbol{\mu}_2\|^2-\|\boldsymbol{\mu}_1\|^2}_{=-b}+\underbrace{2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T}_{=\boldsymbol{w}^T}\boldsymbol{x}&=0\\
\boldsymbol{w}^T\boldsymbol{x}-b&=0
\end{aligned}$$

8 Regression

$$\boldsymbol{w}\triangleq\begin{bmatrix}w_0\\w_1\\w_2\\\vdots\\w_M\end{bmatrix},\qquad\boldsymbol{\phi}(x)\triangleq\begin{bmatrix}1\\x\\x^2\\\vdots\\x^M\end{bmatrix}$$
$$\Rightarrow\hat{f}(x)=\boldsymbol{\phi}^T(x)\cdot\boldsymbol{w}$$

$$\Rightarrow L(\boldsymbol{w})=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{f}(x_i)\right)^2=\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\boldsymbol{\phi}^T(x_i)\,\boldsymbol{w}\right)^2=\frac{1}{N}\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{w}\|_2^2$$

where:

$$\boldsymbol{\Phi}\triangleq\begin{bmatrix}| & & |\\\boldsymbol{\phi}(x_1) & \cdots & \boldsymbol{\phi}(x_N)\\| & & |\end{bmatrix}^T=\begin{bmatrix}1 & x_1 & x_1^2 & \ldots & x_1^M\\\vdots & \vdots & \vdots & \ldots & \vdots\\1 & x_N & x_N^2 & \ldots & x_N^M\end{bmatrix}$$

Thus, the optimal $\boldsymbol{w}$ is given by:

$$\Rightarrow \boxed{\boldsymbol{w} = \left(\boldsymbol{\Phi}^T \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^T \boldsymbol{y} = \boldsymbol{\Phi}^{\dagger} \boldsymbol{y}}$$

9 Linear SVM

The SVM solution satisfies:

$$y_i \left(\langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + b\right) \geq 1$$

Therefore,

$$\Rightarrow \begin{cases} \langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + b \geq 1 & y_i = 1 \\ \langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + b \leq -1 & y_i = -1 \end{cases}$$

Hence:

$$\Rightarrow \begin{cases} \tilde{x}_i \geq 1 & y_i = 1 \\ \tilde{x}_i \leq -1 & y_i = -1 \end{cases}$$

and specifically, the support vectors satisfy:

$$\Rightarrow \begin{cases} \tilde{x}_i = 1 & y_i = 1 \\ \tilde{x}_i = -1 & y_i = -1 \end{cases}, \qquad \text{for all support vector } \tilde{x}_i$$

In words, all the positive samples are on the positive side (of the real line)
and all the negative samples are on the negative side.
Thus, $\tilde{\mathcal{D}}$ is indeed linear separable and the SVM solution is given by:

$$\tilde{w}^* = 1, \qquad \tilde{b} = 0$$

since the problem is centered around the origin.

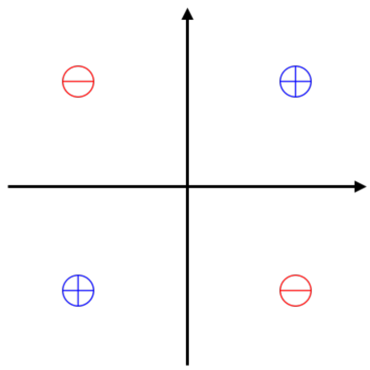
10 Kernels

10.1

$$\begin{aligned} k(x, y) &= (1 + xy)^2 \\ &= 1 + 2xy + x^2 y^2 \\ &= \left\langle \begin{bmatrix} 1 \\ \sqrt{2}x \\ x^2 \end{bmatrix}, \begin{bmatrix} 1 \\ \sqrt{2}y \\ y^2 \end{bmatrix} \right\rangle \end{aligned}$$

$$\Rightarrow \phi(x) = \begin{bmatrix} 1 \\ \sqrt{2}x \\ x^2 \end{bmatrix}$$

10.2



The original data set is:

$$\boldsymbol{X} = \begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} +1 & -1 & +1 & -1 \end{bmatrix}$$

1. This problem is not linear separable with the standard kernel $k(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{x}^T \boldsymbol{z}$.
2. $$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{z}) &= \left(\boldsymbol{x}^T \boldsymbol{z}\right)^2 \\ &= (x_1 z_1 + x_2 z_2)^2 \\ &= x_1^2 z_1^2 + 2 x_1 x_2 z_1 z_2 + x_2^2 z_2^2 \\ &= \left\langle \begin{bmatrix} x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \end{bmatrix}, \begin{bmatrix} z_1^2 \\ \sqrt{2} z_1 z_2 \\ z_2^2 \end{bmatrix} \right\rangle \end{aligned}$$

   Hence, the features in the kernel space are:

   $$\Rightarrow \boldsymbol{\Phi} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} +1 & -1 & +1 & -1 \end{bmatrix}$$

   This can be linearly separated using the second row.
3. The kernel $k(\boldsymbol{x}, \boldsymbol{z}) = \left(1 + \boldsymbol{x}^T \boldsymbol{z}\right)^2$ contains all the features of the kernel $\left(\boldsymbol{x}^T \boldsymbol{z}\right)^2$ and thus, can also be linearly separated.
4. Any data set can be linearly separated with a Gaussian kernel (and some small σ).