Introduction to Machine Learning
Summer 2018
Final Exam Solution

1 ML

$$\begin{aligned}\ell\left(\mu,\sigma^2\right) &= \log\left(\prod_{i=1}^{N} p_X\left(x_i;\mu,\sigma^2\right)\right)\\ &= \sum_{i=1}^{N}\left(\log\left(\frac{1}{x_i}\cdot\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{1}{2\sigma^2}\left(\log\left(x_i\right)-\mu\right)^2\right) \qquad x_i > 0\end{aligned}$$

$$\hat{\mu} = \arg\max_{\mu} \ell\left(\mu,\sigma^2\right)$$

$$\begin{aligned}\frac{\partial}{\partial\mu}\ell\left(\mu,\sigma^2\right) &= 0\\ \sum_{i=1}^{N}\left(\log\left(x_i\right)-\mu\right) &= 0\\ \mu &= \frac{1}{N}\sum_{i=1}^{N}\log\left(x_i\right)\end{aligned}$$

$$\Rightarrow \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\log\left(x_i\right)$$

$$\Rightarrow \exp\left(\hat{\mu}\right) = \exp\left(\frac{1}{N}\sum_{i=1}^{N}\log\left(x_i\right)\right) = \exp\left(\log\left(\left(\prod_{i=1}^{N}x_i\right)^{\frac{1}{N}}\right)\right) = \left(\prod_{i=1}^{N}x_i\right)^{\frac{1}{N}}$$

We obtain the geometric mean.
For $\{x_1 = 2, x_2 = 3\}$ we have:

$$\exp\left(\hat{\mu}\right) = \sqrt{2\cdot 3} = \sqrt{6}$$

2 MAP

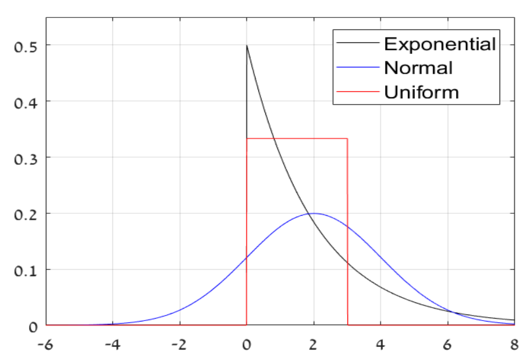
$$\begin{aligned}
\hat{\lambda}_{MAP} &= \arg\max_{\lambda} p\left(\lambda \middle| \{x_i\}\right) \\
&= \arg\max_{\lambda} p\left(\{x_i\} | \lambda\right) p_{\lambda}\left(\lambda\right) \\
&= \arg\max_{\lambda} \prod_{i=1}^{N} \left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda}\right) \cdot \frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)} \lambda^{\alpha-1} e^{-\beta\lambda} \\
&= \arg\max_{\lambda} \prod_{i=1}^{N} \left(\lambda^{x_i} e^{-\lambda}\right) \cdot \lambda^{\alpha-1} e^{-\beta\lambda} \\
&= \arg\max_{\lambda} \lambda^{\sum\limits_{i=1}^{N} x_i + \alpha - 1} e^{-(N+\beta)\lambda} \\
&= \arg\max_{\lambda} \underbrace{\left(\sum_{i=1}^{N} x_i + \alpha - 1\right) \log\left(\lambda\right) - \left(N+\beta\right)\lambda}_{\triangleq f(\lambda)}
\end{aligned}$$

$$\begin{aligned}
f'\left(\lambda\right) &= 0 \\
\left(\sum_{i=1}^{N} x_i + \alpha - 1\right) \frac{1}{\lambda} - N - \beta &= 0 \\
\left(\sum_{i=1}^{N} x_i + \alpha - 1\right) \frac{1}{\lambda} &= N + \beta
\end{aligned}$$

$$\Rightarrow \hat{\lambda}_{MAP} = \frac{\sum_{i=1}^{N} x_i + \alpha - 1}{N+\beta} = \frac{N\overline{x} + \alpha - 1}{N+\beta} \xrightarrow[N\to\infty]{} \overline{x}$$

3 Bayes Classifier

The conditional probabilities densities are as follows



Consider the following cases:

- $x < 0$: In this case, only the conditonal normal distribution has a positive density (thus, the prior distributions are irrelevant).
- $0 \leq x < 3$: The class u and g have the same prior probability and the conditional uniform density is larger than the conditional normal density $\left(\frac{1}{\sqrt{8\pi}} < \frac{1}{3}\right)$. Hence, we need compare between the uniform and the exponential posterior

probabilities

$$
\begin{aligned}
&0.4 \cdot \frac{1}{2} e^{-\frac{x}{2}} \geq 0.3 \cdot \frac{1}{3} \\
\Rightarrow & e^{-\frac{x}{2}} \geq \frac{1}{2} \\
\Rightarrow & -\frac{x}{2} \geq -\ln(2) \\
\Rightarrow & x \leq 2 \ln(2) \approx 1.38.
\end{aligned}
$$

- $x \geq 3$: In this case, the conditional uniform distibution is zero (its prior is irrelvant), hence, we compare the normal and exponential posterior probabilities:

$$
\begin{aligned}
&0.4 \cdot \frac{1}{2} e^{-\frac{x}{2}} \geq 0.3 \cdot \frac{1}{\sqrt{8\pi}} e^{-\frac{(x-2)^2}{8}} \\
\Rightarrow & e^{-\frac{x}{2}} \geq \frac{1.5}{\sqrt{8\pi}} e^{-\frac{(x-2)^2}{8}} \\
\Rightarrow & -\frac{x}{2} \geq \ln(1.5) - \frac{1}{2}\ln(8\pi) - \frac{(x-2)^2}{8} \\
\Rightarrow & x^2 - 4x + 4 - 4x - 8\Big(\ln(1.5) - \frac{1}{2}\ln(8\pi)\Big) \geq 0 \\
\Rightarrow & x^2 - 8x + 4 - 8\Big(\ln(1.5) - \frac{1}{2}\ln(8\pi)\Big) \geq 0 \\
\Rightarrow & x \geq 5.532.
\end{aligned}
$$

Note that the second solution $x \leq 2.468$ is irrelavant since we assumed $x \geq 3$.

Thus, the Bayes optimal classifier is given by

$$
\hat{\omega} = \begin{cases} g & x < 0 \\ e & 0 \leq x < 1.38 \\ u & 1.38 \leq x < 3 \\ g & 3 \leq x < 5.532 \\ e & x \geq 5.532. \end{cases}
$$

4 Histogram

$$
\begin{aligned}
\mathbb{E}\left[\hat{p}_X(x_0)\right] &= \mathbb{E}\left[\frac{1}{N} \cdot \frac{1}{|R_k|} \sum_{i=1}^{N} \boldsymbol{I}\left\{x_i \in R_k\right\}\right] \\
&= \frac{1}{N} \cdot \frac{1}{b-a} \sum_{i=1}^{N} \mathbb{E}\left[\boldsymbol{I}\left\{x_i \in R_k\right\}\right] \\
&= \frac{1}{N} \cdot \frac{1}{b-a} \sum_{i=1}^{N} \Pr\left\{x_i \in R_k\right\} \\
&= \frac{1}{b-a} \Pr\left\{a < x_1 \leq b\right\} \\
&= \frac{F_X(b) - F_X(a)}{b-a}
\end{aligned}
$$

5 PCA I

1.

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

2.

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

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

6 PCA II

As we saw in the tutorial, the new emprical covariance matrix is given

$$
\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_x &= \boldsymbol{V}\boldsymbol{\Sigma}_x\boldsymbol{V}^T \\
&= \boldsymbol{V}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{V}^T \\
&= (\boldsymbol{V}\boldsymbol{U})\boldsymbol{\Lambda}(\boldsymbol{V}\boldsymbol{U})^T \\
&= \tilde{\boldsymbol{U}}\boldsymbol{\Lambda}\tilde{\boldsymbol{U}}^T
\end{aligned}
$$

Therefore, the new priniciple componenets are given by $\tilde{\boldsymbol{U}}=\boldsymbol{V}\boldsymbol{U}$. Denote by $\tilde{\boldsymbol{U}}_m$ the first m priniciple componenets, we have that $\tilde{\boldsymbol{U}}_m=\boldsymbol{V}\boldsymbol{U}_m$, hence

$$
\begin{aligned}
\tilde{\boldsymbol{y}}_i &= \tilde{\boldsymbol{U}}_m^T\tilde{\boldsymbol{x}}_i \\
&= \boldsymbol{U}_m^T\underbrace{\boldsymbol{V}^T\boldsymbol{V}}_{=I}\boldsymbol{x}_i \\
&= \boldsymbol{U}_m^T\boldsymbol{x}_i \\
&= \boldsymbol{y}_i
\end{aligned}
$$

Thus, we can conclude the applying an orthonormal linear transformation does not change the representations.

7 K-Menas

$$
J_0=\sum_{k=1}^{K}\sum_{\boldsymbol{x}_i\in\mathcal{C}_k}\left\|\boldsymbol{x}_i-\boldsymbol{\mu}_k\right\|_2^2
$$

$$J_1 = \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|_2^2$$

where:

$$\boldsymbol{m}_k = \frac{1}{|\mathcal{C}_k|} \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \boldsymbol{x}_i$$

For simplicity, we consider a single cluster:

$$J_0 = \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$$

$$J_1 = \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k\|_2^2$$

$$\begin{aligned}
J_0 &= \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2 \\
&= \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \|\boldsymbol{x}_i - \boldsymbol{m}_k + \boldsymbol{m}_k - \boldsymbol{\mu}_k\|_2^2 \\
&= \sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \left(\|\boldsymbol{x}_i - \boldsymbol{m}_k\|_2^2 + \|\boldsymbol{m}_k - \boldsymbol{\mu}_k\|_2^2 + 2\left(\boldsymbol{x}_i - \boldsymbol{m}_k\right)^T \left(\boldsymbol{m}_k - \boldsymbol{\mu}_k\right)\right) \\
&= J_1 + 2\left(\boldsymbol{m}_k - \boldsymbol{\mu}_k\right)^T \underbrace{\sum_{\boldsymbol{x}_i \in \mathcal{C}_k} \left(\boldsymbol{x}_i - \boldsymbol{m}_k\right)}_{=0} + |\mathcal{C}_k| \, \|\boldsymbol{m}_k - \boldsymbol{\mu}_k\|_2^2 \\
&= J_1 + \underbrace{|\mathcal{C}_k| \, \|\boldsymbol{m}_k - \boldsymbol{\mu}_k\|_2^2}_{\geq 0} \\
&\geq J_1
\end{aligned}$$

$$\Rightarrow J_1 \leq J_0$$

8 Perceptron

8.1

The lower bound is zero (or one) iterations if the initial guess is already providing perfect classification.
For example:

$$\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and:

$$\boldsymbol{w}_1 = \begin{bmatrix} 10 \\ 0 \end{bmatrix}$$

$$\Rightarrow \text{sign}\left(\boldsymbol{w}_1^T \boldsymbol{x}_1\right) = \text{sign}\left(10\right) = 1 = y_1$$

8.2

The upper bound is 11 (or 12 if $\text{sign}(0) \neq 1$) iterations.
For example:

$$\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

and

$$\boldsymbol{w}_1 = \begin{bmatrix} -10 \\ 0 \end{bmatrix}$$

$$\Rightarrow \text{sign}\left(\boldsymbol{w}_1^T \boldsymbol{x}_1\right) = \text{sign}(-10) = -1$$

So after one iteration we have:

$$\boldsymbol{w}_2 = \boldsymbol{w}_1 + \boldsymbol{x}_1 = \begin{bmatrix} -10 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -9 \\ 0 \end{bmatrix}$$

$$\Rightarrow \boldsymbol{w}_3 = \begin{bmatrix} -8 \\ 0 \end{bmatrix}$$

$$\vdots$$

$$\boldsymbol{w}_{11} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\boldsymbol{w}_{12} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\Rightarrow \text{sign}\left(\boldsymbol{w}_{11}^T \boldsymbol{x}_1\right) = \text{sign}(1) = 1 = y_1$$

9 Regression

1.

$$\begin{aligned} \mathrm{d}L &= \left(-\boldsymbol{X}\mathrm{d}\boldsymbol{w}\right)^T \boldsymbol{A}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right) - \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right)^T \boldsymbol{A}\boldsymbol{X}\mathrm{d}\boldsymbol{w} \\ &= -\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right)^T \left(\boldsymbol{A}^T + \boldsymbol{A}\right)\boldsymbol{X}\mathrm{d}\boldsymbol{w} \\ &= \left\langle -\boldsymbol{X}^T\left(\boldsymbol{A}^T + \boldsymbol{A}\right)\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right), \mathrm{d}\boldsymbol{w}\right\rangle \end{aligned}$$

$$\Rightarrow \nabla_{\boldsymbol{w}} L = -\boldsymbol{X}^T\left(\boldsymbol{A}^T + \boldsymbol{A}\right)\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\right)$$

2.

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_0 - \mu\nabla_{\boldsymbol{w}} L = \boldsymbol{w}_0 + \mu\boldsymbol{X}^T\left(\boldsymbol{A}^T + \boldsymbol{A}\right)\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}_0\right)$$

10 Kernel function

$$\begin{aligned} k\left(\boldsymbol{x}, \boldsymbol{z}\right) &= \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{z} \\ &= \boldsymbol{x}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{z} \\ &= \boldsymbol{x}^T\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{z} \\ &= \left\langle \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{x}, \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{z}\right\rangle \end{aligned}$$

$$\Rightarrow \boldsymbol{\phi}\left(\boldsymbol{x}\right) = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{U}^T\boldsymbol{x}$$