JHU Engineering for Professionals
Applied and Computational Mathematics
Data Mining: 625.740 Fall '20

Midterm Exam Solutions

1. The Kullback-Leibler divergence of g from f is

$$D(f,g) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \, \log \left[\frac{f(\mathbf{x})}{g(\mathbf{x})} \right] \, d\mathbf{x}.$$

The distribution g(x) is normal:

$$g(\mathbf{x}) \;\sim\; \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

and

$$g(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}((\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}.$$

Fix the distribution $f(x)$, then

$$\begin{aligned}
D(f,g) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} [f(\mathbf{x}) \, \log f(\mathbf{x}) - f(\mathbf{x}) \, \log g(\mathbf{x})] \; d\mathbf{x} \\
&= \text{const.} + \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \, \left[\frac{1}{2} (\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) + \frac{1}{2} \log |\Sigma| \right] \, d\mathbf{x} \\
0 = \frac{\partial D(f,g)}{\partial \boldsymbol{\mu}} &= - \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \, \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) \, d\mathbf{x}
\end{aligned}$$

Left multiply by Σ so

$$\boldsymbol{\mu} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \, d\mathbf{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathbf{x} \, f(\mathbf{x}) \, d\mathbf{x}$$

$$\boldsymbol{\mu} = \mathbb{E}_{f(\mathbf{x})}[\mathbf{x}],$$

$$\frac{\partial^2 D(f,g)}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^T} = \Sigma^{-1} > 0.$$

The Hessian matrix is positive definite, so we are at a minimum.

$$0 = \frac{\partial D(f,g)}{\partial \, \Sigma^{-1}} = \frac{1}{2} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \, \left[(\mathbf{x}-\boldsymbol{\mu})^T (\mathbf{x}-\boldsymbol{\mu}) - \Sigma \right] \, d\mathbf{x}$$

so

$$\Sigma \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \, d\mathbf{x} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\mathbf{x}) \, (\mathbf{x}-\boldsymbol{\mu})^T (\mathbf{x}-\boldsymbol{\mu}) \, d\mathbf{x}$$

$$\Sigma = \mathbb{E}_{f(\mathbf{x})}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T].$$

Without defining the second derivative of a scalar with respect to a matrix we can still reason that here it is positive by analogy:

$$\frac{\partial}{\partial a^{-1}}(-a) = a^2.$$

$$\frac{\partial}{\partial\,\Sigma^{-1}}\frac{\partial D(f,g)}{\partial\,\Sigma^{-1}} = \frac{\partial}{\partial\,\Sigma^{-1}}(-\Sigma) > 0.$$

Thus the parameters that minimize the Kullback-Liebler divergence are

$$\begin{aligned}\boldsymbol{\mu} &= \mathbb{E}_{f(\mathbf{x})}[\mathbf{x}],\\ \Sigma &= \mathbb{E}_{f(\mathbf{x})}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T].\end{aligned}$$

2.

$$P(\mathbf{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n}\theta_i^{x_i}(1-\theta_i)^{1-x_i},$$

$$\mathcal{D} = \{\mathbf{x}_1,\cdots,\mathbf{x}_k\}$$

(a) $\mathbf{s} = \sum_{j=1}^{k}\mathbf{x}_j = (s_1,\cdots,s_n)^T$

$$\begin{aligned}P(\mathcal{D}|\boldsymbol{\theta}) &= \prod_{j=1}^{k}P(\mathbf{x}_j|\theta),\\ &= \prod_{j=1}^{k}\prod_{i=1}^{n}\theta_i^{x_{i,j}}(1-\theta_i)^{1-x_{i,j}},\\ &= \prod_{i=1}^{n}\prod_{j=1}^{k}\theta_i^{x_{i,j}}(1-\theta_i)^{1-x_{i,j}},\\ &= \prod_{i=1}^{n}\theta_i^{\sum_{j=1}^{k}x_{i,j}}(1-\theta_i)^{\sum_{j=1}^{k}(1-x_{i,j})},\\ &= \prod_{i=1}^{n}\theta_i^{s_i}(1-\theta_i)^{k-s_i},\end{aligned}$$

(b) $\boldsymbol{\theta} \sim \mathcal{U}(0,1)^n$

$$\begin{aligned}P(\boldsymbol{\theta}|\mathcal{D}) &= \frac{P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int_0^1\cdots\int_0^1 P(\mathcal{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}}\\ &= \frac{\prod_{i=1}^{n}\theta_i^{s_i}(1-\theta_i)^{k-s_i}}{\prod_{i=1}^{n}\int_0^1\theta_i^{s_i}(1-\theta_i)^{k-s_i}d\theta_i}\\ &= \frac{\prod_{i=1}^{n}\theta_i^{s_i}(1-\theta_i)^{k-s_i}}{\prod_{i=1}^{n}\frac{s_i!(k-s_i)!}{(k+1)!}}\\ &= \prod_{i=1}^{n}\frac{(k+1)!}{s_i!(k-s_i)!}\theta_i^{s_i}(1-\theta_i)^{k-s_i},\end{aligned}$$

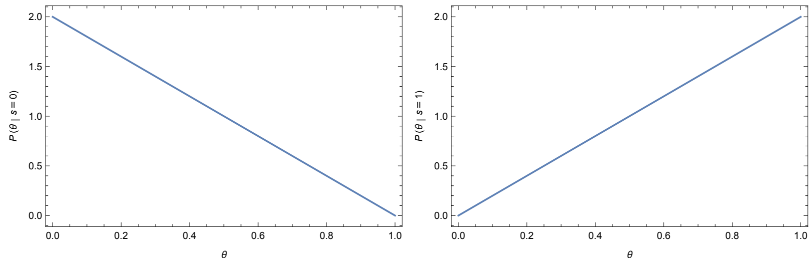

Figure 1: Density $P(\theta|x)$ for $s_1 = 0$ (left) and for $s_1 = 1$ (right).

We have made use of the formula (integrate-by-parts n times)

$$
\begin{aligned}
\int_0^1 \theta^m (1-\theta)^n d\theta &= \frac{1}{m+1} \int_0^1 (1-\theta)^n d\, \theta^{m+1} \\
&= \frac{n}{m+1} \int_0^1 (1-\theta)^{n-1} \theta^{m+1} d\theta \\
&\;\;\vdots \\
&= \frac{m!n!}{(m+n)!} \int_0^1 \theta^{m+n} d\theta \\
&= \frac{m!n!}{(m+n+1)!}
\end{aligned}
$$

(c) Plotted, in Figure 1 are the densities $P(\theta|x)$ for $s = 0$ and $s = 1$.

$$
P(\theta|x) = \begin{cases} 1-\theta, & s = 0, \\ \theta, & s = 1. \end{cases}
$$

(d)

$$
\begin{aligned}
P(x|\mathcal{D}) &= \int_0^1 \cdots \int_0^1 P(\mathbf{x}|\boldsymbol{\theta}) P(\boldsymbol{\theta}|\mathcal{D})\, d\boldsymbol{\theta} \\
&= \prod_{i=1}^n \frac{(k+1)!}{s_i!(k-s_i)!} \int_0^1 \cdots \int_0^1 \theta_i^{s_i+x_i} (1-\theta_i)^{k+1-s_i-x_i} d\boldsymbol{\theta} \\
&= \frac{1}{k+2} \prod_{i=1}^n \frac{(s_i+x_i)!(k+1-s_i-x_i)!}{s_i!(k-s_i)!} \\
&= \prod_{i=1}^n \begin{cases} \left(\dfrac{s_i+1}{k+2}\right)^{x_i}, & x = 1 \\ \left(1 - \dfrac{s_i+1}{k+2}\right)^{1-x_i}, & x = 0 \end{cases} \\
&= \prod_{i=1}^n \left(\frac{s_i+1}{k+2}\right)^{x_i} \left(1 - \frac{s_i+1}{k+2}\right)^{1-x_i}
\end{aligned}
$$

(e) $\hat{\boldsymbol{\theta}} = \dfrac{\mathbf{s}+1}{k+2}$

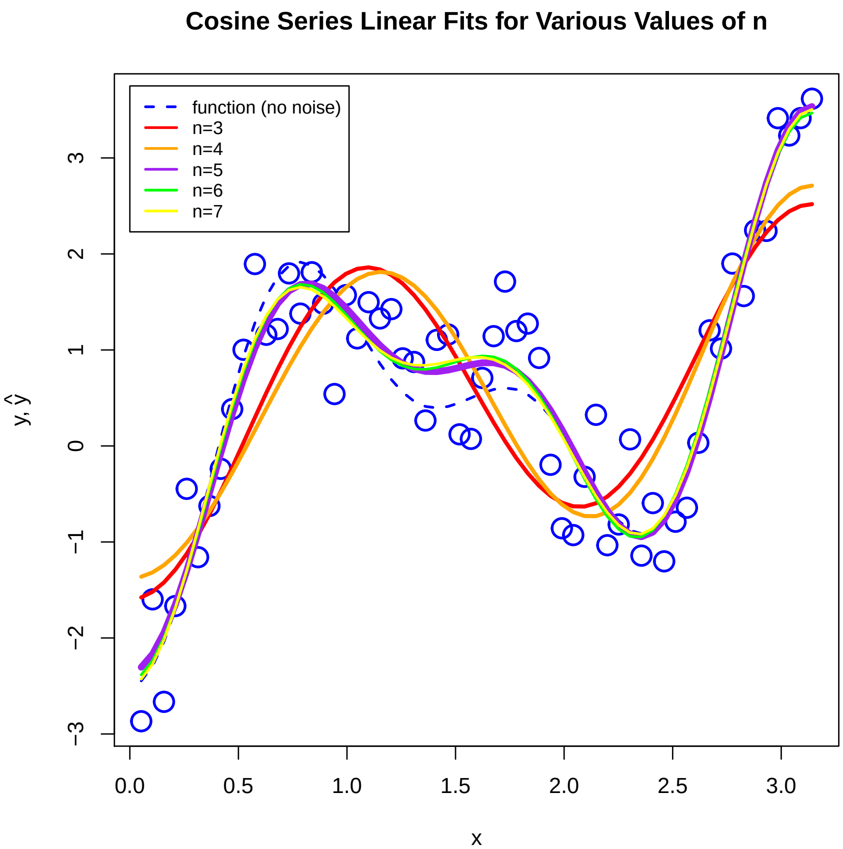

Figure 2: The data in the file `midterm_exam.data.txt` is plotted along with regression curves $y_n = \frac{a_0}{2} + \sum_{k=1}^{n} a_k \cos kx$ for $n = 3, \ldots, 7$.

3. (a) Cosine series fits to the data are shown in Figure 2 for $n = 3, \ldots, 7$. For each $n \in \{5, 6, 7\}$ the series fits the data well. To minimize the complexity of the model among these, we choose $n = 5$.

   (b) The noise can be estimated as $\sqrt{\sum_{j=1}^{N} (\hat{y}_j - y_j)^2 / N} = 0.473$. The noise added to the data was Gaussian with mean $\mu = 0$ and standard deviation $\sigma = 0.500$.