Solutions to Tutorial Week 2

COMP9418 – Advanced Topics in Statistical Machine Learning, 17s2, UNSW Sydney

Instructor: Edwin V. Bonilla

Last Update: Tuesday 1st August, 2017 at 09:56

This tutorial provides a very small sample of problems you should be able to formalize and solve mathematically. If you struggle with the exercises below, I strongly advise you against taking COMP9418 for credits.

1. **Linear Algebra.** Given the matrix $\mathbf{A}$ and column vectors $\mathbf{x}$, $\mathbf{y}$:

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}, \tag{1}$$

   (a) Compute $\mathbf{A}\mathbf{x}$, $\mathbf{x}^T\mathbf{y}$ and $\mathbf{x}\mathbf{y}^T$

   (b) Find all the eigenvalues and eigenvectors of $\mathbf{A}$

Solution

(a) $\mathbf{b} = \mathbf{A}\mathbf{x}$ is a 3-dimensional column vector with components $b_i = \sum_{j=1}^{3} A_{ij}x_j$, hence: $\mathbf{b} = (4, 5, 7)^T$. The scalar $c = \mathbf{x}^T\mathbf{y}$ is an inner product so that $c = \sum_{i=1}^{3} x_i y_i = 5$. $\mathbf{B} = \mathbf{x}\mathbf{y}^T$ is an outer product giving the 3×3 matrix with elements $B_{ij} = x_i y_j$,

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 2 & 1 \\ 6 & 6 & 3 \end{bmatrix}. \tag{2}$$

(b) The eigenvectors of $\mathbf{A}$ satisfy $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ so to obtain the eigenvalues and eigenvectors of $\mathbf{A}$ we write down the characteristic equation:

$$|\mathbf{A} - \lambda\mathbf{I}| = 0, \tag{3}$$

where $|\cdot|$ denotes the determinant. Expanding the above equation we obtain:

$$\left| \begin{bmatrix} 2-\lambda & 1 & 1 \\ 1 & 2-\lambda & 1 \\ 1 & 1 & 2-\lambda \end{bmatrix} \right| = (2-\lambda)^3 + 1 + 1 - (2-\lambda) - (2-\lambda) - (2-\lambda) = 0 \tag{4}$$

$$(1-\lambda)(1-\lambda)(4-\lambda) = 0, \tag{5}$$

where we see that the eigenvalues are $\lambda^{(1)} = 1$ (repeated twice) and $\lambda^{(2)} = 4$. The corresponding eigenvectors are obtained by substituting the eigenvalues in $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ and solving the system of linear equations. For $\lambda = 1$ this places the constraint $v_3 = -(v_1 + v_2)$ and for $\lambda = 4$ this places the constraint $v_1 = v_2 = v_3$ so we have, for example, the eigenvectors $\mathbf{v}^{(1)} = (1, 1, -2)^T$ and $\mathbf{v}^{(2)} = (1, 1, 1)^T$.

2. **Expectation and variance.** Let $X \in \{0, 1\}$ be a Bernoulli random variable, i.e $p(x|\theta) = \theta^x(1-\theta)^{1-x}$. Derive expressions for the expectation $\mathbb{E}[X]$ and variance $\mathbb{V}[X]$. Show all your working.

Solution

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x=0}^{1} x\theta^x(1-\theta)^{1-x} = \theta && (6)\\
\mathbb{V}[X] &= \mathbb{E}\big[(X-\mathbb{E}[X])^2)\big] && (7)\\
&= \sum_{x=0}^{1}(x-\mathbb{E}[X])^2\theta^x(1-\theta)^{1-x} && (8)\\
&= \theta^2(1-\theta)+(1-\theta)^2\theta && (9)\\
&= \theta(1-\theta)(\theta+1-\theta) && (10)\\
&= \theta(1-\theta). && (11)
\end{aligned}$$

3. **The Monty Hall problem.** You have entered a game where there are three boxes, only one of which contains a prize and the other two are empty. Your goal is to pick up the box with the prize in it. You select one of the boxes, and the host of the contest who knows the location of the prize and will not open up that box, opens one of the other boxes and reveals that it is empty. He then gives you to the chance to change your choice. Should you switch to another box? would that increase your chances of winning the prize? Let $C \in \{r, g, b\}$ denote the box that contains the prize where r, g, b refer to the identity of each box.

Solution WLOG assume the following:

- You have selected box r
- Denote the event: "the host opens box b" with H=b

$$p(C=r)=\frac{1}{3} \qquad p(C=g)=\frac{1}{3} \qquad p(C=b)=\frac{1}{3}$$
$$p(H=b|C=r)=\frac{1}{2} \qquad p(H=b|C=g)=1 \qquad p(H=b|C=b)=0$$

We want to compute $p(C=r|H=b)$ and $p(C=g|H=b)$ to decide if we should switch from our initial choice.

We have that:

$$\begin{aligned}
p(H=b) &= \sum_{c\in\{r,g,b\}} p(H=b|C=c)p(C=c)\\
&= (1/2)\,(1/3)+(1)\,(1/3)+(0)\,(1/3)\\
&= 1/2
\end{aligned}$$

Therefore:

$$p(C=r|H=b)=\frac{p(H=b|C=r)p(C=r)}{p(H=b)}=\frac{(1/2)(1/3)}{(1/2)}=1/3$$

Similarly, $p(C=g|H=b)=2/3$.
You should switch from your initial choice to the other box in order to increase your chances of winning the prize!

4. **Unconstrained optimization.** Let $\mathbf{A}$ be a positive definite (PD) symmetric matrix and $\mathbf{x}, \mathbf{b}$ be column vectors. Find the minimum of the function:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}^T\mathbf{x} + c. \qquad (12)$$

Solution Since $\mathbf{A}$ is a PD symmetric matrix we can write:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b} = 0 \tag{13}$$

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}. \tag{14}$$

5. **Constrained optimization.** Let $\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}\}$ be a set of categorical data where $\mathbf{x}$ is a K-dimensional vector encoded using one-hot encoding, i.e. $x_j \in \{0, 1\}$ and $\sum_{k=1}^{K} x_k = 1$. Assume that $\mathbf{x}$ follows a Categorical distribution, i.e. $p(\mathbf{x}) = \text{Cat}(\mathbf{x}|\boldsymbol{\theta}) = \prod_{k=1}^{K} \theta_k^{x_k}$, with $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$, $\theta_k \geq 0$, and $\sum_k^K \theta_k = 1$.

(a) Write down the likelihood of the observations given the model parameters $p(\mathcal{D}|\boldsymbol{\theta})$.

(b) Find the maximum of the data log-likelihood $\mathcal{L} = \log p(\mathcal{D}|\boldsymbol{\theta})$ subject to the constraint $\sum_{k=1}^{K} \theta_k = 1$ and derive maximum likelihood estimates $\widehat{\theta}_{k\text{ML}}$. HINT: Use Lagrange multipliers.

Solution

(a) The likelihood is given by:

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \theta_k^{x_k^{(i)}} = \prod_{k=1}^{K} \theta_k^{m_k}, \tag{15}$$

where $m_k = \sum_{i=1}^{N} x_k^{(i)}$, i.e. the number of times category k appears in the dataset and $\sum_{k=1}^{K} m_k = N$.

(b) We write down the Lagrangian:

$$F(\boldsymbol{\theta}, \lambda) = \sum_{k=1}^{K} m_k \log \theta_k + \lambda(1 - \sum_{j=1}^{K} \theta_j) \tag{16}$$

and solve $\nabla_{\boldsymbol{\theta},\lambda} F(\boldsymbol{\theta}, \lambda) = 0$, which yields the system of equations:

$$\frac{\partial F}{\partial \theta_j} = \frac{m_j}{\theta_j} - \lambda = 0 \rightarrow m_j = \theta_j \lambda,\ j = 1, \dots, K \tag{17}$$

$$\frac{\partial F}{\partial \lambda} = 1 - \sum_{j=1}^{K} \theta_j = 0 \rightarrow \sum_{j=1}^{K} \theta_j = 1. \tag{18}$$

We can solve for λ by summing over j in Equation (17):

$$\sum_{j=1}^{K} m_j = \lambda \sum_{j=1}^{K} \theta_j \tag{19}$$

$$\lambda = N, \tag{20}$$

where we have used Equation (18). Replacing this value back in Equation (17) we obtain:

$$\widehat{\theta}_{j\text{ML}} = \frac{m_j}{N}. \tag{21}$$

6. **Conjugate priors.** Let $\text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) \overset{\text{def}}{=} \frac{1}{\text{B}(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$ denote a Dirichlet distribution with hyperparameters $\boldsymbol{\alpha}$, where $\text{B}(\boldsymbol{\alpha})$ is the normalization constant given by the multivariate Beta function $\text{B}(\boldsymbol{\alpha}) = \int\prod_{k=1}^{K}\theta_k^{\alpha_k-1} = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\alpha_0)}$; $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$ is the Gamma function, which satisfies $\Gamma(1) = 1$ and $\Gamma(x+1) = x\Gamma(x)$; and $\alpha_0 = \sum_{k=1}^{K}\alpha_k$. Show that when using the likelihood $p(\mathcal{D}|\boldsymbol{\theta})$ in item 5a above and a Dirichlet prior $p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$, the posterior distribution is $p(\boldsymbol{\theta}|\mathcal{D},\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}+\mathbf{m})$, where $\mathbf{m} = (m_1,\ldots,m_K)$ and $m_k = \sum_{i=1}^{N}x_k^{(i)}$. Show all your working.

Solution We obtain the posterior distribution by straightforward application of Bayes' rule:

$$p(\boldsymbol{\theta}|\mathcal{D},\boldsymbol{\alpha}) = \frac{p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\mathcal{D}|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\mathcal{D}|\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{22}$$

$$= \frac{\frac{1}{\text{B}(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}\prod_{k=1}^{K}\theta_k^{m_k}}{\int\frac{1}{\text{B}(\boldsymbol{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}\prod_{k=1}^{K}\theta_k^{m_k}d\boldsymbol{\theta}} \tag{23}$$

$$= \frac{\prod_{k=1}^{K}\theta_k^{m_k+\alpha_k-1}}{\int\prod_{k=1}^{K}\theta_k^{m_k+\alpha_k-1}d\boldsymbol{\theta}}, \tag{24}$$

where we immediately recognize the denominator as the normalization constant of a Dirichlet with parameters $\boldsymbol{\alpha}+\mathbf{m}$, i.e. $\int\prod_{k=1}^{K}\theta_k^{m_k+\alpha_k-1}d\boldsymbol{\theta} = \text{B}(\boldsymbol{\alpha}+\mathbf{m}) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k+m_k)}{\Gamma(\alpha_0+m_0)}$, where $m_0 = \sum_k m_k$. Then:

$$p(\boldsymbol{\theta}|\mathcal{D},\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}+\mathbf{m}). \tag{25}$$