DS-GA.1013 Mathematical Tools for Data Science :
Homework Assignment 0
Yves Greatti - yg390

1. Projections

   (a) False Consider $\boldsymbol{b_1} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ and $\boldsymbol{b_2} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, they form a basis of $\mathbf{R}^2$. When using the definition $\mathcal{P}_{\mathcal{S}} x = \sum_{i=1}^n \langle x, b_i \rangle b_i$ we would expect that $\mathcal{P}_{\mathcal{S}} b_1 = b_1$. However $\mathcal{P}_{\mathcal{S}} b_1 = \begin{bmatrix} 2 \\ 5 \end{bmatrix} \neq b_1$.

   (b) True Let $S^\perp = \{x | \langle x, y \rangle = 0, \forall y \in S\}$ a subspace of an inner product space X, then $S^{\perp\perp} = \{x | \langle x, y \rangle = 0, \forall y \in S^\perp\}$. The inner product being symmetric, $S \subseteq S^{\perp\perp}$. Since for any vector $x \in X$, we have $x = y + z$ where $y \in S, z \in S^\perp$, using Gram-schmidt orthonormalization process, we can find a basis of S and $S^\perp$ which express any vector of X as a linear combination of these two basis and combining these two basis together forms a new basis for X so $\dim X = \dim S + \dim S^\perp$. If $\dim X = n$ and $\dim S = m$ then $\dim S^\perp = n - m$. Similarly $\dim S^{\perp\perp} = n - (n - m) = m$ so $\dim S^{\perp\perp} = \dim S$, so $S^{\perp\perp} \subseteq S$ and since the dimension of a space or subspace is the cardinality of its basis, thus $S = S^{\perp\perp}$.

   (c) True consider $\boldsymbol{v} = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$, we want $\boldsymbol{w} = \begin{bmatrix} \frac{\sum_{i=1,n} v_i}{n} \\ \vdots \\ \frac{\sum_{i=1,n} v_i}{n} \end{bmatrix}$. The orthogonal projection of $\boldsymbol{v}$ onto the vector $\boldsymbol{b}$ is defined as $\frac{v.b}{\|b\|^2}$, take $b = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$.

2. Scalar linear approximation

   (a) First we write $\mathrm{E}[(ax + b - y)^2] = \mathrm{E}[((ax - y) - (-b))^2]$, we know that the best mean-squared error mimimizer of a random variable is its mean so $-b = \mathrm{E}[ax - y] = a\,\mathrm{E}[x] - \mathrm{E}[y] = a\mu_x - \mu_y$. Substituting b in the expression we want to minimize gives us:

$$\begin{aligned}
\mathrm{E}[(ax + b - y)^2] &= \mathrm{E}[(ax - y - (a\mu_x - \mu_y))^2] \\
&= \mathrm{E}[\{a(\mu_x - x) - (y - \mu_y)\}^2] \\
&= a^2\,\mathrm{E}[(x - \mu_x)^2] + \mathrm{E}[(y - \mu_y)^2] - 2a\,\mathrm{E}[(x - \mu_x)(y - \mu_y)] \\
&= a^2 \sigma_x^2 + \sigma_y^2 - 2\,a\,\mathrm{Cov}(x, y)
\end{aligned}$$

   Let $f(a) = a^2\sigma_x^2 + \sigma_y^2 - 2\,a\,\mathrm{Cov}(x, y)$, then $f'(a) = 2(\sigma_x^2 a - \mathrm{Cov}(x, y))$ and $f''(a) = 2\sigma_x^2$. The function is strictly convex, and its second derivative is positive, thus its minimizer is $a = \frac{\mathrm{Cov}(x,y)}{\sigma_x^2} = \rho_{x,y}\,\frac{\sigma_y}{\sigma_x}$.

3. Gradients

   (a) Compute the gradient of $f(x) = b^T x$ where $b \in \mathbf{R}^d$ and $f : \mathbf{R}^d \to \mathbf{R}$. $\frac{\partial f(x)}{x_j} = \sum_i b_i \frac{\partial x_i}{\partial x_j} = b_i$, thus $\nabla f(x) = b$.

   (b) Compute the gradient of $f(x) = x^T A x$ where $A \in \mathbf{R}^{d \times s}$ and $f : \mathbf{R}^d \to \mathbf{R}$. $f(x) = x^T A x = \sum_{i=1}^d \sum_{j=1}^d a_{ij} x_i x_j$, then

$$
\begin{aligned}
\frac{\partial f}{\partial x_k} &= \sum_{i=1}^d \sum_{j=1}^d a_{ij} \frac{\partial x_i x_j}{x_k} \\
&= \sum_{i=1}^d \sum_{j=1}^d a_{ij} (x_j \delta_{ik} + x_i \delta_{jk}) \\
&= \sum_{i=1}^d \sum_{j=1}^d a_{ij} x_j \delta_{ik} + \sum_{i=1}^d \sum_{j=1}^d a_{ij} x_i \delta_{jk} \\
&= \sum_{j=1}^d a_{kj} x_j + \sum_{i=1}^d a_{ik} x_i \\
&= (Ax)_k + (Ax)_k^T
\end{aligned}
$$

   thus $\nabla f(x) = (A + A^T)x$.