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

1. (Projections)Are the following statements true or false? Prove that they are true or provide a counterexample.

   (a) The projection of a vector on a subspace $\mathcal{S}$ is equal to

   $$\mathcal{P}_{\mathcal{S}}\, x = \sum_{i=1}^{n} \langle x, b_i \rangle\, b_i$$

   for any basis $b_1, \ldots, b_d$ of $\mathcal{S}$. 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) The orthogonal complement of the orthogonal complement of a subspace $\mathcal{S} \subseteq \mathbb{R}^n$ is $\mathcal{S}$. 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) Replacing each entry of a vector in $\mathbb{R}^n$ by the average of all its entries is equivalent to projecting the vector onto a subspace. 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. (Eigen decomposition) The populations of deer and wolfs in Yellowstone are well approximated by

$$d_{n+1} = \frac{5}{4} d_n - \frac{3}{4} w_n, \tag{1}$$

$$w_{n+1} = \frac{1}{4} d_n + \frac{1}{4} w_n, \qquad n = 0, 1, 2, \ldots, \tag{2}$$

where d_n and w_n denote the number of deer and wolfs in year n. Assuming that there are more deer than wolfs to start with ($w_0 < d_0$), what is the proportion between the numbers of deer and wolfs as $n \to \infty$?

Rewriting the problem in a matrix form:

$$\begin{pmatrix} d_{n+1} \\ w_{n+1} \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 5 & -3 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} d_n \\ w_n \end{pmatrix}$$

Let $A = \frac{1}{4} \begin{pmatrix} 5 & -3 \\ 1 & 1 \end{pmatrix}$, $v_{n+1} = \begin{pmatrix} d_{n+1} \\ w_{n+1} \end{pmatrix}$, $v_0 = \begin{pmatrix} d_0 \\ w_0 \end{pmatrix}$ then $v_{n+1} = Av_n = A^n v_0$. We are looking to find the eigen decomposition so we can understand the behavior of v_n as $n \to \infty$. $\det(A - \lambda I) = \frac{1}{2}(2\lambda^2 - 3\lambda + 1)$, we find for eigenvalues $\lambda_1 = \frac{1}{2}$ and $\lambda_2 = 1$ with corresponding eigenvectors

$w_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, w_2 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$. Since A is diagonalizable the vectors $\{w_1, w_2\}$ forms a basis of $\mathbf{R}^2$ and we can express v_0 in this basis as $v_0 = \alpha w_1 + \beta w_2$ for some $\alpha, \beta \in \mathbf{R}$, thus $v_{n+1} = \alpha A^n w_1 + \beta A^n w_2 = \alpha \lambda_1^n w_1 + \beta \lambda_2^n w_2 = \alpha(\frac{1}{2^n})w_1 + \beta w_2$. Then taking the $n \to \infty$, the first term goes to zero and $v_{n+1} \sim \beta w_2$. So asymptotically $\frac{d_{n+1}}{w_{n+1}} \sim 3$ which verifies the initial condition: $w_0 < d_0$.

3. Function approximation) In this problem we will work in the real inner product space $L^2[-1,1]$ given by

$$L^2[-1,1] = \left\{ f : [-1,1] \to \mathbb{R} \;\middle|\; \int_{-1}^{1} f(x)^2 \, dx < \infty \right\}.$$

On this space, the inner product is given by

$$\langle f, g \rangle = \int_{-1}^{1} f(x) g(x) \, dx.$$

In the following exercises, you may use a computer to perform the integral calculations.

(a) The functions $\{1, x, x^2\}$ form a basis for the 3-dimensional subspace P_2 of $L^2[-1,1]$ consisting of the polynomials of degree at most 2. Give the orthonormal basis for P_2 obtained by applying Gram-Schmidt to this set of functions.

Using Gram-Schmidt orthonormalization process, we find

$$\begin{aligned} v_1 &= 1 \\ v_2 &= x - \langle x, 1 \rangle \frac{1}{\langle 1, 1 \rangle} \\ &= x \\ v_3 &= x^2 - \langle x^2, v_2 \rangle \frac{v_2}{\langle v_2, v_2 \rangle} - \langle x^2, v_1 \rangle \frac{v_1}{\langle v_1, v_1 \rangle} \\ &= x^2 - \frac{1}{3} \end{aligned}$$

Then we normalize each of these vectors to obtain:

$$w_1 = \frac{v_1}{\|v_1\|} = \frac{\sqrt{2}}{2}$$
$$w_2 = \frac{v_2}{\|v_2\|} = \sqrt{\frac{3}{2}}\ x$$
$$w_3 = \frac{v_3}{\|v_3\|} = \sqrt{\frac{45}{8}}\ (x^2 - \frac{1}{3})$$

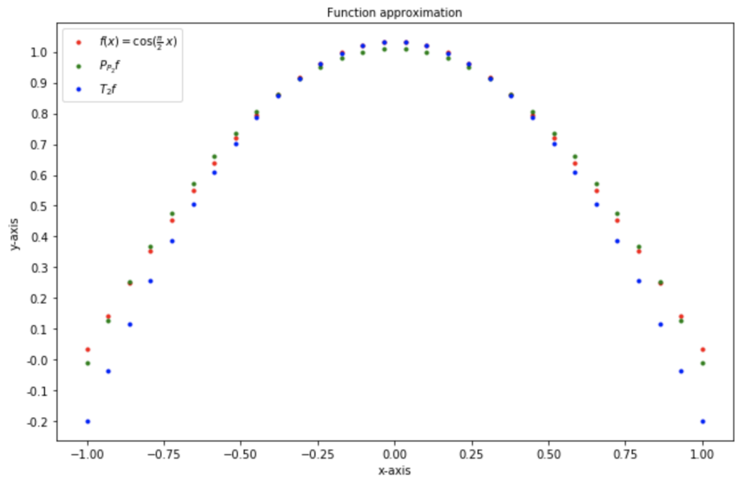
(b) Compute the orthogonal projection of $f(x) = \cos(\pi x/2)$ onto P_2.
The projection of $f(x) = \cos(\frac{\pi}{2}\ x)$ in the orthonormal basis $\{w_1, w_2, w_3\}$ is: $\sum_{i=1,3} \langle f, w_i \rangle w_i$, where:

$$\begin{aligned}
\langle f, w_1 \rangle &= \int_{-1}^{1} \cos(\frac{\pi}{2}\ x)\ \frac{\sqrt{2}}{2}\ dx \\
&= \frac{4}{\pi\sqrt{2}} \sim 0.9 \\
\langle f, w_2 \rangle &= \int_{-1}^{1} \cos(\frac{\pi}{2}\ x)\ \frac{\sqrt{3}}{2}\ x\ dx \\
&= 0 \\
\langle f, w_3 \rangle &= \int_{-1}^{1} \cos(\frac{\pi}{2}\ x) \sqrt{\frac{45}{8}}\ (x^2 - \frac{1}{3})\ dx \\
&= 2\sqrt{10}\frac{\pi^2 - 12}{\pi^3} \sim -0.43
\end{aligned}$$

(c) Plot $f(x) = \cos(\pi x/2)$, $\mathcal{P}_{P_2} f$, and $T_2 f$ on the same axis. Here $\mathcal{P}_{P_2} f$ is the projection computed in the previous part, and $T_2 f$ is the quadratic Taylor polynomial for f centered at $x = 0$:

$$T_2 f(x) = f(0) + f'(0)x + \frac{f''(0)}{2}x^2.$$

Include this plot in your submitted homework document.



(d) The plot from the previous part shows that $\mathcal{P}_{P_2} f$ is a better approximation than $T_2 f$ over most of $[-1, 1]$. Explain why this is the case.

$\mathcal{P}_{P2} f$is the orthogonal projection of $f(x)$ over the subspace of polynomials of degree 2: $\{w_1, w_2, w_3\}$, like the Taylor expansion $\mathcal{T}_2 f$. The difference is that the Taylor polynomial is a polynomial expansion of f at 0. So in a neighborhood of 0, there is almost no differences between f and $\mathcal{T}_2 f$, but as we move away the approximation given by $\mathcal{T}_2 f$ is worst than $\mathcal{P}_{P2} f$.

4. (Scalar linear estimation)

(a) Let $\tilde{x}$ be a random variable with mean $\mu_{\tilde{x}}$ and variance $\sigma_{\tilde{x}}^2$, and $\tilde{y}$ a random variable with mean $\mu_{\tilde{y}}$ and variance $\sigma_{\tilde{y}}^2$. The correlation coefficient between them is $\rho_{\tilde{x},\tilde{y}}$. What values of a, $b \in \mathbb{R}$ minimize the mean square error $\mathrm{E}[(a\tilde{x} + b - \tilde{y})^2]$? Express your answer in terms of $\mu_{\tilde{x}}$, $\sigma_{\tilde{x}}$, $\mu_{\tilde{y}}$, $\sigma_{\tilde{y}}$, and $\rho_{\tilde{x},\tilde{y}}$.

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:

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

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

(b) Let $\tilde{x} = \tilde{y}\tilde{z}$, where $\tilde{y}$ has mean $\mu_{\tilde{y}}$ and variance $\sigma^2_{\tilde{y}}$, and $\tilde{z}$ has mean zero and variance $\sigma^2_{\tilde{z}}$. If $\tilde{y}$ and $\tilde{z}$ are independent, what is the best linear estimate of $\tilde{y}$ given $\tilde{x}$?

Applying the result from the previous question, the best linear estimate of y given x is $y = \rho_{x,y}\frac{\sigma_y}{\sigma_x}(x - \mu_x) + \mu_y$. Notice that $\mathrm{Var}(x) = \mathrm{Var}(y\ z) = \mathrm{E}[y^2\ z^2] - \mathrm{E}[y\ z]^2 = \mathrm{E}[y^2]\mathrm{E}[z^2] - \mathrm{E}[y]^2 E[z]^2 = (\sigma_y^2 + \mu_y^2)\sigma_z^2$ where we have used that a and z are independent and z has zero-mean. And $\mathrm{E}[x] = \mathrm{E}[y\ z] = \mathrm{E}[y]\ .\ 0 = 0$. Thus the best linear estimate of y given x is: $\rho_{x,y}\frac{\sigma_y}{\sigma_z\sqrt{\sigma_y^2+\mu_y^2}}\ x + \mu_y$.

(c) Assume $\tilde{y}$ is positive with probability one. Can you think of a zero-mean random variable $\tilde{z}$ such that $\tilde{y}$ can be estimated perfectly from $\tilde{x}$ in the previous question? If $\tilde{z}$ is a random variable taking the values -1, +1 with equal probability ($\frac{1}{2}$ each) then $\tilde{y}$ can be estimated perfectly from $\tilde{x}$. If $\tilde{z}$ has value -1, then since $\tilde{y}$ is positive with probability one, $\tilde{x} \leq 0$ and $\tilde{y}$ has the opposite value of $\tilde{x}$, when $\tilde{z}$ has value +1, $\tilde{y}$ has the same value as $\tilde{x}$. Also notice that $\mathrm{E}[\tilde{z}] = \frac{1}{2}\ (-1) + \frac{1}{2}\ (+1) = 0$.

5. (Gradients) Recall that the entries of the gradient of a function are equal to its partial derivatives. Use this fact to:

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

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