Machine Learning, Spring 2018: Exercise Sheet 7 – solutions

It's paper and pencil time again.

Problem 1 (A very toy-ish demo of PCA) Assume you have a sample S of four 2-dimensional datapoints from $\mathbb{R}^2$, $S = \{(1,1)', (0,0)', (0,0)', (-1, -1)'\}$. What are the two principal component vectors $\mathbf{u}_1$, $\mathbf{u}_2$ of this dataset?

Solution. The first PC vector is $\mathbf{u}_1 = (1,1)' / \| (1,1)' \| = (1,1)' / \sqrt{2}$. The second PC vector is a unit-length vector orthogonal to $\mathbf{u}_1$, that is $\mathbf{u}_2 = (1, -1)' / \sqrt{2}$. Note: the negatives of $\mathbf{u}_1$, $\mathbf{u}_2$ would likewise qualify as PC vectors, because PC vectors are unique only up to their sign.

Problem 2 (understanding an extreme case of linear regression). Consider a slightly weird dataset $(x_i, y_i)_{i=1,\dots,N} = (2, y_i)_{i=1,\dots,N}$ where all argument-value pairs have the same argument $x_i = 2$ and $y_i \in \mathbb{R}$. If you carry out a linear regression, what is the regression weight vector w that you get from this dataset? It is rather easy to guess the answer – can you also *prove* it?

Solution. The regression weight "vector" is the scalar (1-dimensional vector) $w = \mu / 2$, where μ is the mean of the y_i. To prove this one has to show that the mean μ of the y_i minimizes the average square distance to the data points, that is

$$\mu = \operatorname{argmin}_{\upsilon \in \mathbb{R}} \sum_i (\upsilon - y_i)^2 .$$

To show this we differentiate the sum with respect to ν:

$$\frac{d}{d\upsilon} \sum_i (\upsilon - y_i)^2 = 2 \sum_i (\upsilon - y_i)$$

This becomes zero for $\nu = \mu$.

Problem 3 (understanding the mathematical nature of a combined PCA-feature-extraction + linear regression learning procedure) In the last paragraph of Section 6 in the lecture notes, I state that $D: \mathbb{R}^n \to \mathbb{R}^k$ is an affine map. That is, D can be written as $D(x) = M x + b$, where M is a $k \times n$ sized matrix and b is a k-dimensional vector. Work out the details – write down M and b as constructed from the n-dimensional mean pattern vector μ, the PC matrix U_m, and the regression weight matrix W.

Solution. We proceed in three steps. In the first step we describe the data centering as an affine map. Centering means to subtract from raw n-dimensional data points x the mean μ. This is obviously achieved by the affine map

$$C: \mathbb{R}^n \to \mathbb{R}^n, x \mapsto I_n x - \mu,$$

where I_n is the n-dimensional identity matrix. In the second step we reduce the dimension to m by projection on the first m PCs, giving the feature map

$\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^m, x \mapsto U_m' (Cx) = U_m' I_n x - U_m' \mu = U_m' x - U_m' \mu,$

which is likewise an affine map. In the third step we apply the linear regression weight matrix W, giving us

$D: \mathbb{R}^n \to \mathbb{R}^k, x \mapsto W \mathbf{f}(x) = W U_m' x - W U_m' \mu,$

again an affine map, with $M = W U_m'$ and $b = - W U_m' \mu$.