Notes on notation: basics

- Data are usually given as $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ are called *data points* and $y_i \in \mathbb{R}$ are called *labels*.
- We then assemble the $\boldsymbol{x}_i$ into a *data matrix* $X \in \mathbb{R}^{d \times n}$, where the columns of X are the $\boldsymbol{x}_i$:

$$X = \begin{pmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{pmatrix},$$

and the labels into a label vector $\boldsymbol{y} = (y_1, \dots, y_n) \in \mathbb{R}^{1 \times n}$.

Notes on notation: weight vector and random variable

- Given a data matrix $X \in \mathbb{R}^{d \times n}$ and a (weight-) vector $\boldsymbol{w} \in \mathbb{R}^{d}$ we sometimes calculate $\boldsymbol{w}^\top X$

$$\boldsymbol{w}^\top X = \left(w^\top \boldsymbol{x}_1, \dots, w^\top \boldsymbol{x}_n \right),$$

  thus, $\boldsymbol{w}^\top X$ is a row vector. Therefore, we sometimes write $\boldsymbol{w}^\top X \in \mathbb{R}^{1 \times n}$, to emphasize that it is a row.
- We denote the matrix whose columns are all equal to the mean of all data points as

$$\bar{X} = (\bar{\boldsymbol{x}}, \dots, \bar{\boldsymbol{x}}), \text{ with } \bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i.$$

- We denote a random variable as $X : \Omega \to \mathbb{R}$. Unfortunately, this coincides with the notation for the data matrix! We are using the same symbol for two different objects here.

Notes on notation: feature maps in linear regression

- Maps $\phi : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ or $\varphi : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$ are often called *feature maps*, because they formalize the idea that a data point $\boldsymbol{x}_i$ gives rise to features $\varphi(\boldsymbol{x}_i)$ in a different space. Note that ϕ and φ are just variants of the same letter! In the context of linear regression, maps ϕ are also called *basis functions*.
- Basic example

$$\phi : \mathbb{R}^d \to \mathbb{R}^{d+1}$$

$$\phi(\boldsymbol{x}) = \begin{pmatrix} 1 \\ x_1 \\ \dots \\ x_d \end{pmatrix}$$

Note that here x_i denotes the components of $\boldsymbol{x}$. For $\boldsymbol{w} \in \mathbb{R}^{d+1}$, we have $\boldsymbol{w}^\top \phi(\boldsymbol{x}) = w_0 + w_1 x_1 + \ldots + w_n x_n$.

Notes on notation: feature maps

- Another example

$$\phi : \mathbb{R} \to \mathbb{R}^3$$
$$\phi(x) = \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix}$$

For $\boldsymbol{w} \in \mathbb{R}^3$, we have $\boldsymbol{w}^\top \phi(x) = w_0 + w_1 x + w_2 x^2$. This allows using polynomial basis functions in linear regression. Note that the components of $\boldsymbol{w}$ still appear *linearly*!

Notes on notation

- We sometimes write $(1, \boldsymbol{w})^\top$ to express $(1, w_1, \ldots, w_d)^\top$

Notes on notation: kernels and feature maps

- In the context of kernels, we write $\boldsymbol{x} \in \mathcal{X}$, $\varphi : \mathcal{X} \to \mathcal{F}$, where $\mathcal{F}$ is some (usually high-dimensional) space with a scalar product.
- To emphasize the use of a scalar product, we write $\varphi(\boldsymbol{x})^\top \cdot \varphi(\boldsymbol{x}')$ (For purists: it would be more correct to write either $\varphi(\boldsymbol{x})^\top \varphi(\boldsymbol{x}')$ or $\varphi(\boldsymbol{x}) \cdot \varphi(\boldsymbol{x}')$, because $\cdot$ already denotes the scalar product, hence no need to transpose.)
- Given a data matrix X with columns $\boldsymbol{x}_i$, we also use the notation $\varphi(X)$ (for example in the derivation of kernel ridge regression). This notation is not self-evident, it is defined as follows, for $\varphi : \mathbb{R}^d \to \mathbb{R}^{\tilde{d}}$:

$$\varphi(X) = \begin{pmatrix} \varphi(\boldsymbol{x}_1) & \dots & \varphi(\boldsymbol{x}_n) \end{pmatrix} = \begin{pmatrix} \varphi_1(\boldsymbol{x}_1) & \dots & \varphi_1(\boldsymbol{x}_n) \\ & \dots & \\ \varphi_{\tilde{d}}(\boldsymbol{x}_1) & \dots & \varphi_{\tilde{d}}(\boldsymbol{x}_n) \end{pmatrix},$$

where $\varphi_1, \dots, \varphi_{\tilde{d}}$ denote the components of the feature map φ.

Notes on notation: feature maps continued

We have $\boldsymbol{x} \in \mathbb{R}^d$, $\varphi(\boldsymbol{x}) \in \mathbb{R}^{\tilde{d}}$, $X \in \mathbb{R}^{d \times n}$, $\varphi(X) \in \mathbb{R}^{\tilde{d} \times n}$, $\varphi(X)^\top \in \mathbb{R}^{n \times \tilde{d}}$.

$$
\begin{aligned}
\varphi(X)\varphi(X)^\top &= \begin{pmatrix} \varphi_1(\boldsymbol{x}_1) & \dots & \varphi_1(\boldsymbol{x}_n) \\ & \dots & \\ \varphi_{\tilde{d}}(\boldsymbol{x}_1) & \dots & \varphi_{\tilde{d}}(\boldsymbol{x}_n) \end{pmatrix} \begin{pmatrix} \varphi_1(\boldsymbol{x}_1) & \dots & \varphi_{\tilde{d}}(\boldsymbol{x}_1) \\ & \dots & \\ \varphi_1(\boldsymbol{x}_n) & \dots & \varphi_{\tilde{d}}(\boldsymbol{x}_n) \end{pmatrix} \\
&= \begin{pmatrix} \sum_i \varphi_1(\boldsymbol{x}_i)\varphi_1(\boldsymbol{x}_i) & \dots & \sum_i \varphi_1(\boldsymbol{x}_i)\varphi_{\tilde{d}}(\boldsymbol{x}_i) \\ & \dots & \\ \sum_i \varphi_{\tilde{d}}(\boldsymbol{x}_i)\varphi_1(\boldsymbol{x}_i) & \dots & \sum_i \varphi_{\tilde{d}}(\boldsymbol{x}_i)\varphi_{\tilde{d}}(\boldsymbol{x}_i) \end{pmatrix}
\end{aligned}
$$

This expression occurs in the derivation of ridge regression.

Notes on notation: feature maps continued

$$
\begin{aligned}
\varphi(X)^\top \varphi(X) &= \begin{pmatrix} \varphi_1(\boldsymbol{x}_1) & \dots & \varphi_{\tilde{d}}(\boldsymbol{x}_1) \\ & \dots & \\ \varphi_1(\boldsymbol{x}_n) & \dots & \varphi_{\tilde{d}}(\boldsymbol{x}_n) \end{pmatrix} \begin{pmatrix} \varphi_1(\boldsymbol{x}_1) & \dots & \varphi_1(\boldsymbol{x}_n) \\ & \dots & \\ \varphi_{\tilde{d}}(\boldsymbol{x}_1) & \dots & \varphi_{\tilde{d}}(\boldsymbol{x}_n) \end{pmatrix} \\
&= \begin{pmatrix} \varphi(\boldsymbol{x}_1)^\top \varphi(\boldsymbol{x}_1) & \dots & \varphi(\boldsymbol{x}_1)^\top \varphi(\boldsymbol{x}_n) \\ & \dots & \\ \varphi(\boldsymbol{x}_n)^\top \varphi(\boldsymbol{x}_1) & \dots & \varphi(\boldsymbol{x}_n)^\top \varphi(\boldsymbol{x}_n) \end{pmatrix} \\
&= \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ & \dots & \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \dots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{pmatrix}
\end{aligned}
$$

This expression occurs in the derivation of kernel ridge regression.