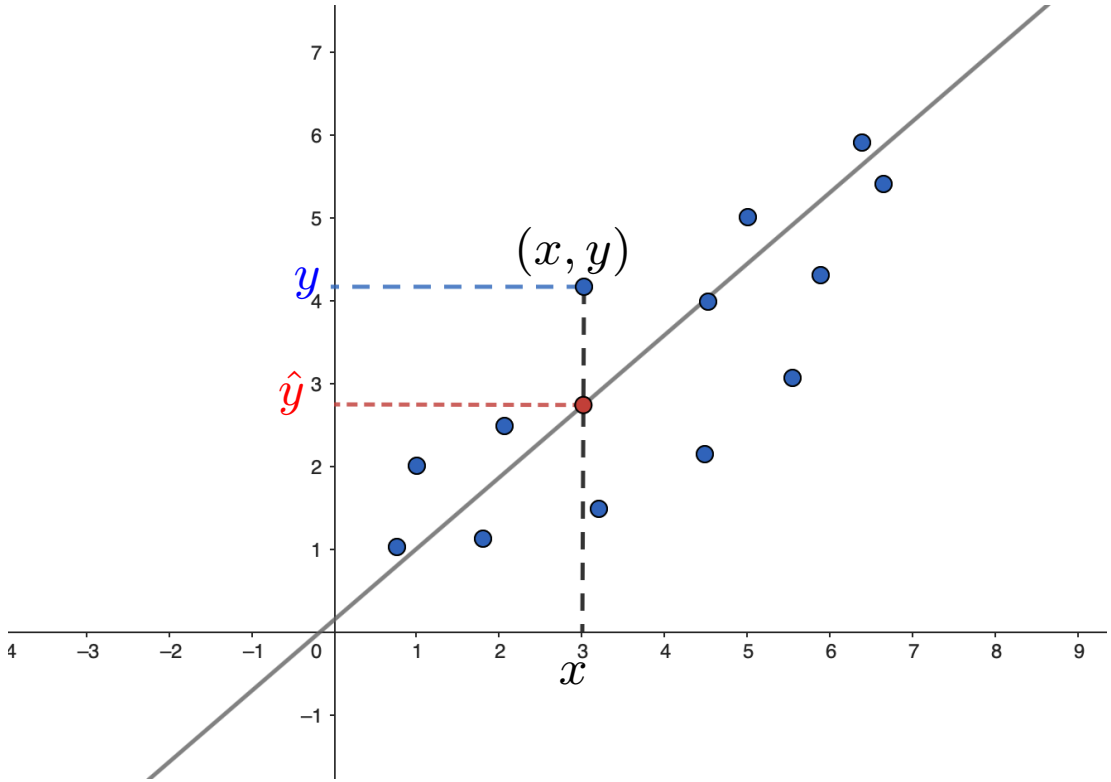
Regression

Regression

- With regression with denote the task that use input data to predict a continuous (real) label
- According to the function we will use to predict the labels we will talk of linear or non-linear regression



Regression model

- Given a set of elements $\mathcal{X} \subseteq \mathbb{R}^n$ and the corresponding labels set $\mathcal{Y} \subseteq \mathbb{R}$, we want to find a function $f : \mathbb{R}^n \to \mathbb{R}$ such that

$$f(x_i) \simeq y_i \quad \forall x_i \in \mathcal{X}$$

- Since the labels are real numbers we cannot ask for a perfect prediction like in the classification problem. Instead, we will look for a predicted label that is "close" to the true one
- We will use loss functions that measure the error between the true labels and the predicted ones:

$$\begin{aligned} L : & \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+ \\ & (y, y') \mapsto |y - y'|^2 \end{aligned}$$

Linear Regression

- The simplest regression model is the linear regression, where we will look for the hyper-plane that best approximate the points $\mathcal{X} \times \mathcal{Y}$
- Given a kernel function $\Phi : \mathcal{X} \to \mathbb{R}^N$ we consider the following hypothesis set for the linear regression:

$$\mathcal{H} = \{x \mapsto w \cdot \Phi(x) + b | w \in \mathbb{R}^N, b \in \mathbb{R}\}$$

- The best hypothesis is chosen based on the quadratic empirical error, i.e. we want the hyper-plane (w, b) that is a solution for the problem:

$$\min_{w,b} \frac{1}{m} \sum_{i=1}^{m} (w \cdot \Phi(x_i) + b - y_i)^2$$

where (x_i, y_i) are the input points

Linear Regression

- It is possible to reformulate the linear regression problem as

$$\min_W F(W) = \frac{1}{m}\|X^\top W - Y\|^2$$

  dove
  - $X = \begin{pmatrix} \Phi(x_1) & \cdots & \Phi(x_m) \\ 1 & \cdots & 1 \end{pmatrix}$
  - $W = \begin{pmatrix} w_1 & \cdots & w_m & b \end{pmatrix}^\top$
  - $Y = \begin{pmatrix} y_1 & \cdots & y_m \end{pmatrix}^\top$
- Since the function F is convex, it admits a global minimum in W if and only if $\nabla F(W) = 0$:

$$\frac{2}{m} X(X^\top W - Y) = 0 \Leftrightarrow XX^\top W = XY$$

Linear Regression

- Since the function F is convex, it admits a global minimum in W if and only if $\nabla F(W) = 0$:

$$\frac{2}{m} X(X^\top W - Y) = 0 \Leftrightarrow XX^\top W = XY$$

- If $XX^\top$ is invertible the problem admits a unique solution

$$W = (XX^\top)^{-1} XY$$

- If $XX^\top$ is not invertible there exists a family of solutions:

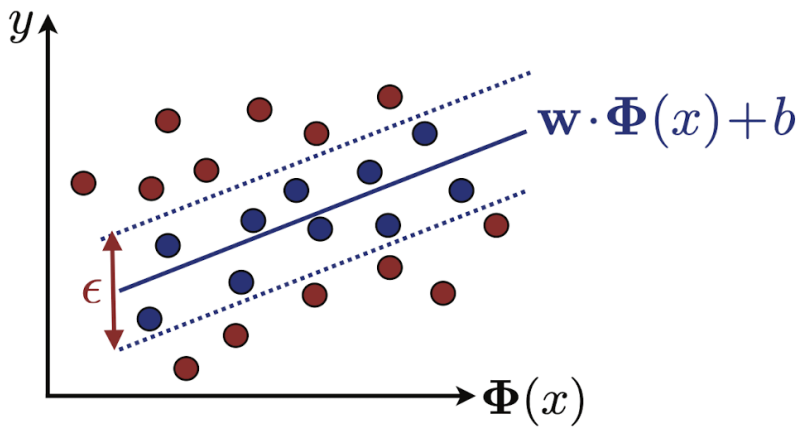
$$W = (XX^\top)^\dagger XY + (\mathrm{Id} - (XX^\top)^\dagger XX^\top) W_0$$

  where W_0 is any matrix in $\mathbb{R}^{N \times N}$ and $(XX^\top)^\dagger$ is the **Moore-Penrose pseudo-inverse** of $XX^\top$

- Usually we consider the element in the family with the least norm, i.e. $W = (XX^\top)^\dagger XY$

Support Vector Regression

- The regression model based on support vectors is inspired by SVM
- The idea is to find a tubular region of radius ε that best approximate the input data



- We obtain two class of points
  - Points that lie inside the tubular region and so they have at most distance ε to the prediction function
  - Points outside the tubular region and that will be penalized in terms of their distance to the prediction function

Support Vector Regression

- Given a hypothesis family $\mathcal{H} = \{x \mapsto w \cdot \Phi(x) + b\}$ with Φ being a function kernel, we can formulate the SVR problem as

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}|y_i - w \cdot \Phi(x_i) + b|_\varepsilon$$

where

$$|y - y'|_\varepsilon = \max\{0, |y - y'| - \varepsilon\}$$

- By introducing the slack variables $\xi, \xi' \geq 0$ we obtain the equivalent problem

$$\begin{cases} \min\limits_{w,b} & \frac{1}{2}\|w\|^2 + C\sum\limits_{i=1}^{m}(\xi_i + \xi_i') \\ s.t. & w \cdot \Phi(x_i) + b - y_i \leq \varepsilon + \xi_i \\ & y_i - (w \cdot \Phi(x_i) + b) \leq \varepsilon + \xi_i' \\ & \xi_i, \xi_i' \geq 0 \end{cases}$$

Support Vector Regression

- The previous problem is a quadratic convex problem, therefore it is possible to obtain its dual using the KKT conditions

$$\begin{cases} \min\limits_{\alpha,\alpha'} & -\varepsilon(\alpha'+\alpha)^\top \mathbf{1} + (\alpha'-\alpha)^\top \mathbf{y} - \frac{1}{2}(\alpha'-\alpha)^\top \mathbf{K}(\alpha'-\alpha) \\ s.t. & 0 \leq \alpha \leq C \\ & 0 \leq \alpha' \leq C \\ & (\alpha'-\alpha)^\top \mathbf{1} = 0 \end{cases}$$

- The prediction function can be computed by a dual solution α, α' as

$$f(x) = \sum_{i=1}^{m} (\alpha_i' - \alpha_i) K(x_i, x) + b$$

- where the term b can be obtained in two ways
  - if x_j is such that $0 < \alpha_j < C$:

$$b = -\sum_{i=1}^{m} (\alpha_i' - \alpha_i) K(x_i, x_j) + y_j + \varepsilon$$

  - if x_j is such that $0 < \alpha_j' < C$:

$$b = -\sum_{i=1}^{m} (\alpha_i' - \alpha_i) K(x_i, x_j) + y_j - \varepsilon$$

Singular Value Decomposition (SVD)

- Given a matrix $A \in \mathbb{R}^{m \times n}$ with rank $r \leq \min\{m, n\}$ we can define the following quantities
  - $\sigma_1, \ldots \sigma_r$, given by the square root of the eigenvalues of di $A^\top A$;
  - $\Sigma \in \mathbb{R}^{m \times n}$ rectangular diagonal with $\Sigma_{i,i} = \sigma_i$;
  - $U \in \mathbb{R}^{m \times m}$ whose columns are eigenvectors of $AA^\top$;
  - $V \in \mathbb{R}^{n \times n}$ whose columns are eigenvectors of $A^\top A$
- The SVD decomposition of A is the factorization

$$A = U\Sigma V^\top$$

Singular Value Decomposition

We can use the SVD of a matrix A in several ways

- the Moore-Penrose pseudo-inverse of A is given by

$$A^{\dagger} = U\Sigma^{\dagger}V^{\top}$$

  where $\Sigma^{\dagger}$ is a rectangular diagonal matrix with $\Sigma^{\dagger}_{i,i} = \sigma_i^{-1}$

- Given a matrix $A \in \mathbb{R}^{m \times n}$ with rank r, if we want to find the matrix $X^* \in \mathbb{R}^{m \times n}$ with rank $k \leq r$ that best approximate A, i.e.

$$X^* = \arg\min_X \|A - X\|_F$$

  The Eckart-Young Theorem proves that $X^* = U\tilde{\Sigma}V^{\top}$, where $\tilde{\Sigma}$ is equal to Σ but all the singular values σ_i with $i > k$ are zeroed.

back