Mathematics for Machine Learning

— Empirical Risk Minimization

Joseph Chuang-Chieh Lin

Department of Computer Science & Information Engineering,
Tamkang University

Fall 2023

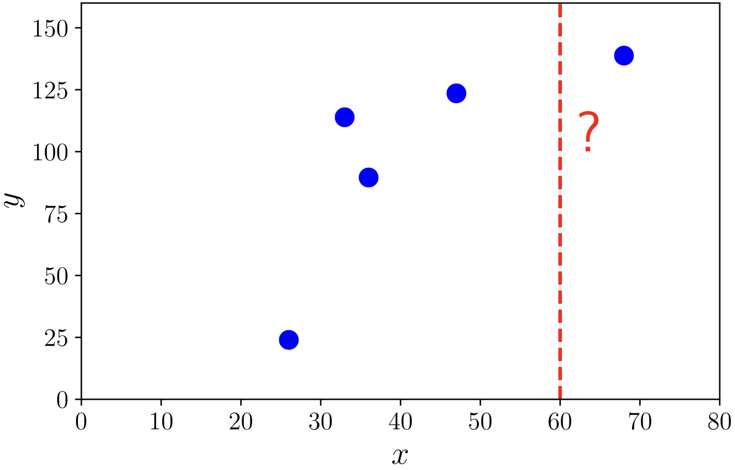
Credits for the resource

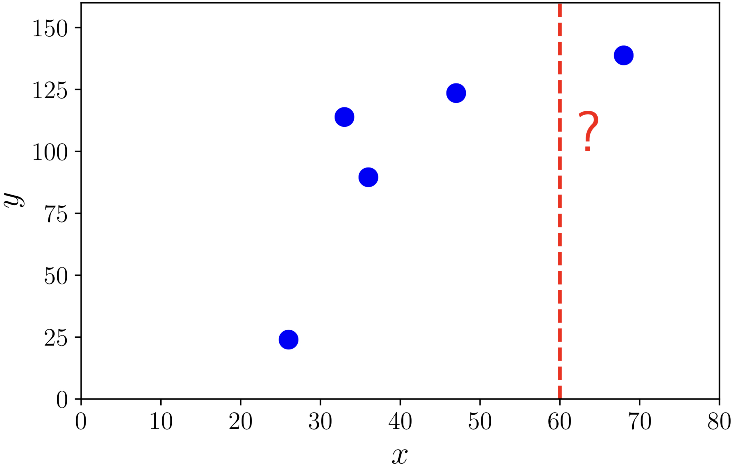
- The slides are based on the textbooks:
  - *Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong: Mathematics for Machine Learning. Cambridge University Press. 2020.*
  - *Howard Anton, Chris Rorres, Anton Kaul: Elementary Linear Algebra. Wiley. 2019.*
- We could partially refer to the monograph:
  Francesco Orabona: A Modern Introduction to Online Learning. https://arxiv.org/abs/1912.13213

Outline

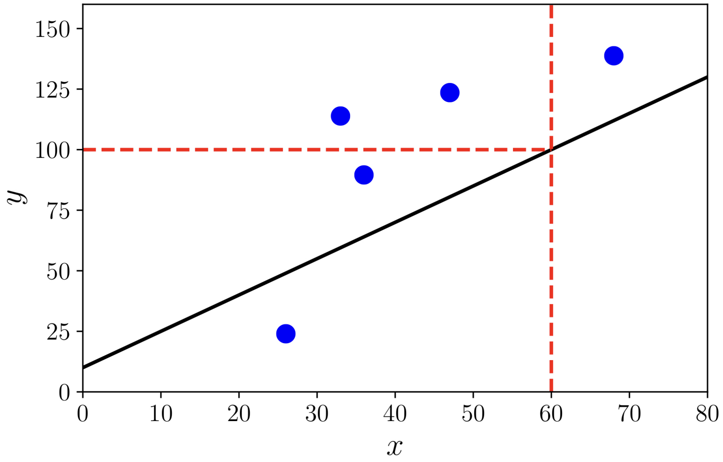
Motivation

- It's time to consider a problem that a ML algorithm is designed to solve.
- We will see some performance metrics to speak for what a "good" model is.
- As before, we assume that the data is represented as vectors.
- Denote by N the number of examples (or data points, examples, etc.) in a dataset.
- The data has D features, hence a vector is of D-dimensional here.


150
125
100
75
50
25
0
y
0
10
20
30
40
50
60
70
80
x
?



- We are interested in the salary of a person aged 60.



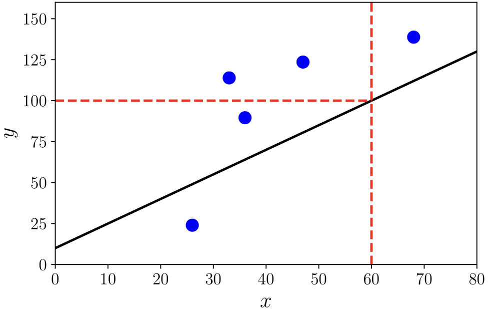
- We are interested in the salary of a person aged 60.

Models as Functions

For example, consider the linear function $f\colon \mathbb{R}^D \mapsto \mathbb{R}$,

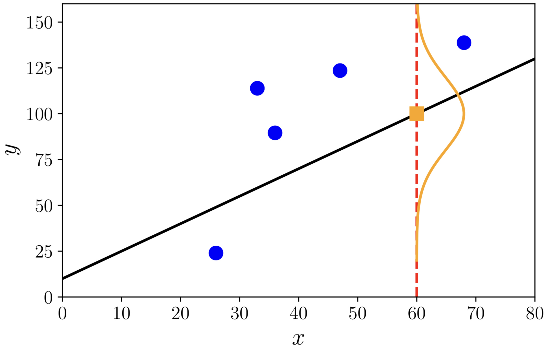
$$f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} + \theta_0$$

for unknown $\boldsymbol{\theta}$ and θ_0.



Models as Probability Distributions

We can also consider predictors as probabilistic models (e.g., distribution of possible functions).



Goal of Learning

- Find a model and its corresponding parameters such that the predictor performs well on unseen data.
- Three algorithmic phases:
  - Prediction or inference
    - Non-probabilistic: prediction (e.g., Empirical risk minimization (ERM)).
    - Probabilistic: inference (e.g., maximum likelihood, Bayesian inference).
  - Training or parameter estimation.
  - Hyperparameter tuning or model selection.

Outline

Hypothesis Class of Functions

Given N examples $\mathbf{x}_i \in \mathbb{R}^D$, $i = 1, \ldots, N$ and corresponding labels $y_i \in \mathbb{R}$.

Goal: Estimate a predictor $f(\cdot, \boldsymbol{\theta}) : \mathbb{R}^D \mapsto \mathbb{R}$, parametrized by $\boldsymbol{\theta}$

$$f(\mathbf{x}_i, \boldsymbol{\theta}^{*}) \approx y_i \ \text{ for all } i \in \{1, \ldots, N\},$$

where $\boldsymbol{\theta}^*$ is a good parameter we aim to find.

Let $\hat{y}_i = f(\mathbf{x}_i, \boldsymbol{\theta}^*)$ represent the output of the predictor.

Example

Consider the set of affine functions.

- Let $\mathbf{x}_i = [1, x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(D)}]^\top$
- The corresponding parameter $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_D]^\top$.
- Consider a more compact form as below:

$$f(\mathbf{x}_i, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}_i.$$

Example

Consider the set of affine functions.

- Let $\mathbf{x}_i = [1, x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(D)}]^\top$
- The corresponding parameter $\boldsymbol{\theta} = [\theta_0, \theta_1, \ldots, \theta_D]^\top$.
- Consider a more compact form as below:

$$f(\mathbf{x}_i, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}_i.$$

which is equivalent to

$$f(\mathbf{x}_i, \boldsymbol{\theta}) = \theta_0 + \sum_{d=1}^{D} \theta_d x_i^{(d)}$$

Loss Functions for Training & Empirical Risk

We specify a loss function $\ell(y_n, \hat{y}_n)$ to say how *bad* a model fits the data.

Goal: Loss Minimization

Find a good parameter $\boldsymbol{\theta}^*$ such that the average loss on the set of N training examples is minimized.

Assumptions

A given training set $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$ is independently and identically distributed (i.i.d.).

- $\boldsymbol{X} := [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times D}$, label vector $\mathbf{y} := [y_1, \ldots, y_N]^\top \in \mathbb{R}^N$.
- The average loss:

$$\boldsymbol{R}_{\text{emp}}(f, \boldsymbol{X}, \mathbf{y}) = \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, \hat{y}_i).$$

Example

Consider the squared loss $\ell(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$. So we wish to solve

Example

Consider the squared loss $\ell(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$. So we wish to solve

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^D} \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_n, \boldsymbol{\theta}))^2,$$

that is,

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^D} \frac{1}{N} \sum_{i=1}^{N} (y_i - \boldsymbol{\theta}^\top \mathbf{x}_i)^2$$

Example

Consider the squared loss $\ell(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2$. So we wish to solve

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^D} \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_n, \boldsymbol{\theta}))^2,$$

that is,

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^D} \frac{1}{N} \sum_{i=1}^{N} (y_i - \boldsymbol{\theta}^\top \mathbf{x}_i)^2 \iff \min_{\boldsymbol{\theta} \in \mathbb{R}^D} \frac{1}{N} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2.$$

⋆ The least-squares problem.

Remark: True Risk in Terms of Expected Risk (1/2)

- We are NOT interested in a predictor that ONLY performs well on the training data.
- We seek a predictor that performs well on unseen test data.

Remark: True Risk in Terms of Expected Risk (1/2)

- We are NOT interested in a predictor that ONLY performs well on the training data.
- We seek a predictor that performs well on unseen test data.
- Formally, we are interested in finding f that minimizes the expected risk:

$$\mathbf{R}_{\text{true}}(f) = \mathbb{E}_{\mathbf{x},y}[\ell(y, f(\mathbf{x}))],$$

where y is the label and $f(\mathbf{x})$ is the prediction based on $\mathbf{x}$.

  - ⋆ $\mathbf{R}_{\text{true}}(f)$: the true risk if we had access to an infinite amount of data.

Remark: True Risk in Terms of Expected Risk (2/2)

Questions arising from minimizing expected risk:

- How should we change the training procedure to generalize well?
- How do we estimate expected risk from finite data?

Regularization: An Approach to Reduce Overfitting

Key: Bias the search for the minimizer of empirical risk by introducing a penalty term which is referred to as regularization.

Example

Revisit the least-squares problem. By adding a penalty term involving $\boldsymbol{\theta}$ we have:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^D} \frac{1}{N} \|\mathbf{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|^2.$$

Cross-Validation: Assess the Generalization Performance (1/2)

Partition the dataset into two sets $\mathcal{D} = \mathcal{R} \cup \mathcal{V}$ s.t. $\mathcal{R} \cap \mathcal{V} = \emptyset$.

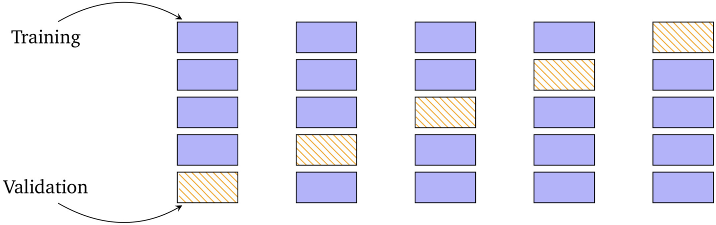
- $\mathcal{R}$: the training set.
- $\mathcal{V}$: the validation set.

Cross-Validation: Assess the Generalization Performance (1/2)

Partition the dataset into two sets $\mathcal{D} = \mathcal{R} \cup \mathcal{V}$ s.t. $\mathcal{R} \cap \mathcal{V} = \emptyset$.

- $\mathcal{R}$: the training set.
- $\mathcal{V}$: the validation set.

K-fold cross-validation: partition the data into K chunks ($K - 1$ of them: $\mathcal{R}$; the rest one of them: $\mathcal{V}$).



Cross-Validation: Assess the Generalization Performance (1/2)

Cross-validation approximates the expected generalization error:

$$\mathbb{E}_{\mathcal{V}}[R(f, \mathcal{V})] \approx \frac{1}{K} \sum_{k=1}^{K} R(f^{(k)}, \mathcal{V}^{(k)}),$$

where $R(f^{(k)}, \mathcal{V}^{(k)})$ is the risk (e.g., RMSE) on the validation set $\mathcal{V}^{(k)}$ for predictor $f^{(k)}$.

- A potential computational cost of training the model K times, which can be burdensome (except we can do it in parallel).

Discussions