Introduction to Machine Learning

Chapter 0: Notation and definitions

Bernd Bischl, Christoph Molnar, Daniel Schalk, Fabian Scheipl

Department of Statistics – LMU Munich

FUNDAMENTAL DEFINITIONS AND NOTATION

- $\mathcal{X}$: p-dim. **input space**, usually we assume $\mathcal{X} = \mathbb{R}^p$, but categorical **features** can occur as well.
- $\mathcal{Y}$: **target space**, e. g. $\mathcal{Y} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$, $\mathcal{Y} = \{-1, 1\}$, $\mathcal{Y} = \{1, \dots, g\}$ or $\mathcal{Y} = \{\text{label}_1 \dots \text{label}_g\}$
- x : **feature vector**, $x = (x_1, \dots, x_p)^T \in \mathcal{X}$
- y : target / label / output. $y \in \mathcal{Y}$
- $\mathbb{P}_{xy}$: joint probability distribution on $\mathcal{X} \times \mathcal{Y}$
- $p(x, y)$ or $p(x, y|\theta)$: joint pdf for x and y

FUNDAMENTAL DEFINITIONS AND NOTATION

Remark:

This lecture is mainly developed from a frequentist perspective. If parameters appear behind the |, this is for better reading, and does not imply that we condition on them in a Bayesian sense (but this notation would actually make a Bayesian treatment simple). So formally, $p(x|\theta)$ should be read as $p_\theta(x)$ or $p(x, \theta)$ or $p(x; \theta)$.

FUNDAMENTAL DEFINITIONS AND NOTATION

- $\left(x^{(i)}, y^{(i)}\right)$: i-th **observation** or **instance**
- $\mathcal{D} = \left\{\left(x^{(1)}, y^{(1)}\right), \ldots, \left(x^{(n)}, y^{(n)}\right)\right\}$: **data set** with n observations
- $\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}$: data for training and testing, often $\mathcal{D} = \mathcal{D}_{\text{train}} \dot{\cup}\, \mathcal{D}_{\text{test}}$
- $f(x)$ or $f(x|\theta) \in \mathbb{R}$ or $\mathbb{R}^g$: prediction function (**model**) learned from data, we might suppress θ in notation
- $h(x)$ or $h(x|\theta) \in \mathcal{Y}$: discrete prediction for classification (see later)
- $\theta \in \Theta$: model **parameters**
  (some models may traditionally use different symbols)
- H : hypothesis space. f lives here, restricts the functional form of f
- $\epsilon = y - f(x)$ or $\epsilon^{(i)} = y^{(i)} - f\left(x^{(i)}\right)$: **residual** in regression
- $yf(x)$ or $y^{(i)} f\left(x^{(i)}\right)$: **margin** for binary classification with $\mathcal{Y} = \{-1, 1\}$ (see later)

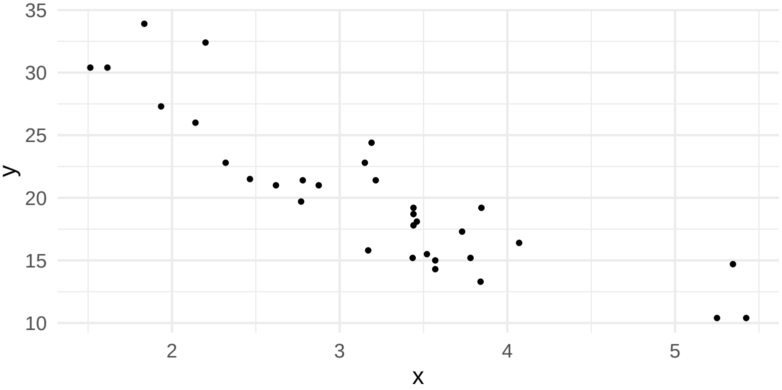
FUNDAMENTAL DEFINITIONS AND NOTATION

- $\pi_k(x) = \mathbb{P}(y = k|x)$: posterior probability for class k, given x, in case of binary labels we might abbreviate $\pi(x) = \mathbb{P}(y = 1|x)$
- $\pi_k = \mathbb{P}(y = k)$: prior probability for class k, in case of binary labels we might abbreviate $\pi = \mathbb{P}(y = 1)$
- $\mathcal{L}(\theta)$ and $\ell(\theta)$: Likelihood and log-Likelihood for a parameter θ, based on a statistical model
- $\hat{f}$, $\hat{h}$, $\hat{\pi}_k(x)$, $\hat{\pi}(x)$ and $\hat{\theta}$: learned functions and parameters

Remark: With a slight abuse of notation we write random variables, e.g., x and y, in lowercase, as normal variables or function arguments. The context will make clear what is meant.

FUNDAMENTAL DEFINITIONS AND NOTATION

In the simplest case we have i.i.d. data $\mathcal{D}$, where the input and output space are both real-valued and one-dimensional.



FUNDAMENTAL DEFINITIONS AND NOTATION

Design matrix (with or w/o intercept term):

$$X = \begin{pmatrix} x_1^{(1)} & \cdots & x_p^{(1)} \\ \vdots & \vdots & \vdots \\ x_1^{(n)} & \cdots & x_p^{(n)} \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_1^{(1)} & \cdots & x_p^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1^{(n)} & \cdots & x_p^{(n)} \end{pmatrix}$$

- $\mathbf{x}_j = \left(x_j^{(1)}, \ldots, x_j^{(n)}\right)^T$: j-th observed feature vector.
- $\mathbf{y} = \left(y^{(1)}, \ldots, y^{(n)}\right)^T$: vector of target values.
- The right design matrix demonstrates the trick to encode the intercept via an additional constant-1 feature, so the feature space will be $(p+1)$-dimensional. This allows to simplify notation, e.g., to write $f(x) = \theta^T x$, instead of $f(x) = \theta^T x + \theta_0$.

BINARY LABEL CODING

Remark: Notation in binary classification can be sometimes confusing because of different coding styles, and as we have to talk about predicted scores, classes and probabilities.

A binary variable can take only two possible values. For probability / likelihood-based model derivations a 0-1-coding, for geometric / loss-based models the -1/+1-coding is often preferred:

- $\mathcal{Y} = \{0, 1\}$. Here, the approach often models $\pi(x)$, the posterior probability for class 1 given x. Usually, we then define $h(x) = [\pi(x) \geq 0.5] \in \mathcal{Y}$.
- $\mathcal{Y} = \{-1, 1\}$. Here, the approach often models $f(x)$, a real-valued score from $\mathbb{R}$ given x. Usually, we define $h(x) = \text{sign}(f(x)) \in \mathcal{Y}$, and we interpret |f(x)| as “confidence” for the predicted class $h(x)$.