I2ML :: Basics

Notation

$\mathcal{X}$: p-dim. **input space**

Usually we assume $\mathcal{X} = \mathbb{R}^p$, but categorical **features** can also occur

$\mathcal{Y}$: **target space**

E.g.: $\mathcal{Y} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$, $\mathcal{Y} = \{-1, 1\}$, $\mathcal{Y} = \{1, \ldots, g\}$

$\mathrm{x} = (x_1, \ldots, x_p)^T \in \mathcal{X}$: **feature vector**

$y \in \mathcal{Y}$: **target / label / output**

$\mathbb{P}_{xy}$: **Joint probability distribution on** $\mathcal{X} \times \mathcal{Y}$

$p(\mathrm{x}, y)$ or $p(\mathrm{x}, y \mid \boldsymbol{\theta})$: **joint probability density function (pdf)**

$\left(\mathrm{x}^{(i)}, y^{(i)}\right)$: i -th **observation** or **instance**

$\mathcal{D} = \left(\left(\mathrm{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathrm{x}^{(n)}, y^{(n)}\right)\right)$

data set with n observations.

$\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{test}}$: data for training and testing

Often, $\mathcal{D} = \mathcal{D}_{\text{train}} \dot{\cup} \mathcal{D}_{\text{test}}$.

$f(\mathrm{x})$ or $f(\mathrm{x} \mid \boldsymbol{\theta}) \in \mathbb{R}$ or $\mathbb{R}^g$: prediction function (**model**)

We might suppress $\boldsymbol{\theta}$ in notation.

$h(\mathrm{x})$ or $h(\mathrm{x}|\boldsymbol{\theta}) \in \mathcal{Y}$

Discrete prediction for classification.

$\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_d) \in \Theta$: model **parameters**

Some models may traditionally use different symbols.

$\mathcal{H}$: **hypothesis space**

f lives here, restricts the functional form of f.

$\epsilon = y - f(\mathrm{x})$ or $\epsilon^{(i)} = y^{(i)} - f\left(\mathrm{x}^{(i)}\right)$

Residual in regression.

$yf(\mathrm{x})$ or $y^{(i)} f\left(\mathrm{x}^{(i)}\right)$: **margin** for binary classification

With, $\mathcal{Y} = \{-1, 1\}$.

$\pi_k(\mathrm{x}) = \mathbb{P}(y = k \mid \mathrm{x})$: **posterior probability** for class k, given x

In case of binary labels we might abbreviate $\pi(\mathrm{x}) = \mathbb{P}(y = 1 \mid \mathrm{x})$.

$\pi_k = \mathbb{P}(y = k)$: **prior probability** for class k

In case of binary labels we might abbreviate $\pi = \mathbb{P}(y = 1)$.

$\mathcal{L}(\boldsymbol{\theta})$ and $\ell(\boldsymbol{\theta})$: Likelihood and log-Likelihood for a parameter $\boldsymbol{\theta}$

These are based on a statistical model.

$\hat{y}$, $\hat{f}$, $\hat{h}$, $\hat{\pi}_k(\mathrm{x})$, $\hat{\pi}(\mathrm{x})$ and $\hat{\boldsymbol{\theta}}$

These are learned functions and parameters (These are estimators of corresponding functions and parameters).

Concepts

Model: $f : \mathcal{X} \to \mathbb{R}^g$ is a function that maps feature vectors to predictions.

Learner: takes a data set with features and outputs (**training set**, $\in \mathcal{X} \times \mathcal{Y}$) and produces a **model** (which is a function $f : \mathcal{X} \to \mathbb{R}^g$)

Learning = Representation + Evaluation + Optimization.

Representation: (Hypothesis space) Defines which kind of model structure of f can be learned from the data.

Example: Linear functions, Decision trees etc.

Evaluation: A metric that quantifies how well a specific model performs on a given data set. Allows us to rank candidate models in order to choose the best one.

Example: Squared error, Likelihood etc.

Optimization: Efficiently searches the hypothesis space for good models.

Example: Gradient descent, Quadratic programming etc.

Loss function: The "goodness" of a prediction $f(\mathrm{x})$ is measured by a loss function $L(y, f(\mathrm{x}))$ Through **loss**, we calculate the prediction error and the choice of the loss has a major influence on the final model

Risk Minimization: The ability of a model f to reproduce the association

between x and y that is present in the data $\mathcal{D}$ can be measured by the average loss: the **empirical risk**.

$$\mathcal{R}_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathrm{x}^{(i)}\right)\right).$$

Learning then amounts to **empirical risk minimization** – figuring out which model f has the smallest average loss:

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f).$$

Regression Losses

Basic Idea (L2 loss/ squared error):

$L(y, f(\mathrm{x})) = (y - f(\mathrm{x}))^2$ or $L(y, f(\mathrm{x})) = 0.5(y - f(\mathrm{x}))^2$

Convex and differentiable.

Tries to reduce large residuals (if residual is twice as large, loss is 4 times as large)

Basic Idea (L1 loss/ absolute error):

$L(y, f(\mathrm{x})) = |y - f(\mathrm{x})|$

Convex and more robust

No derivatives for $= 0$, $y = f(\mathrm{x})$, optimization becomes harder

$\hat{f}(\mathrm{x}) =$ median of $y|\boldsymbol{x}$

Classification

We want to assign new observations to known categories according to criteria learned from a training set.

Assume we are given a **classification problem:**

$x \in \mathcal{X}$	feature vector
$y \in \mathcal{Y} = \{1, \ldots, g\}$	*categorical* output variable (label)
$\mathcal{D} = \left(\left(\mathrm{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathrm{x}^{(n)}, y^{(n)}\right)\right)$	observations of x and y

Classification usually means to construct g discriminant functions:

$f_1(\mathrm{x}), \ldots, f_g(\mathrm{x})$, so that we choose our class as $h(\mathrm{x}) = \arg\max_k f_k(\mathrm{x})$ for $k = 1, 2, \ldots, g$

Linear Classifier:

If the functions $f_k(\mathrm{x})$ can be specified as linear functions, we will call the classifier a *linear classifier*.

Note: All linear classifiers can represent non-linear decision boundaries in our original input space if we include derived features. For example: higher order interactions, polynomials or other transformations of x in the model.

Binary classification: If only 2 classes exist

We can use a single discriminant function $f(\mathrm{x}) = f_1(\mathrm{x}) - f_2(\mathrm{x})$.