Machine Learing

Matteo Secco

March 7, 2021

Contents

1 Supervised Learning

input x also called features or attributes

output t also called targets or labels

Goal find a good approximation for $f : x \to t$

1.1 Tasks

Classification t is discrete

Regression t is continuous

Probability estimation t is a probability

1.2 When to use supervised learning?

- Human cannot perform the task
- Human can perform the task but cannot explain how
- Task changes over time
- Task is user-specific

1.3 Steps

To approximate f over the dataset $\mathcal{D}$

1. Define a loss function $\mathcal{L}$
2. Choose the hypothesis space $\mathcal{H}$
3. find $h \in \mathcal{H}$ that minimizes $\mathcal{L}$ over $\mathcal{D}$

1.4 Representation, Evaluation, Optimization

Examples of representation

- Linear models
- Instance-based
- Decision trees
- Set of rules
- Graphical models
- Neural networks

- Gaussian Processes
- Support vector machines
- Model ensembles

Examples of evaluation

- Accuracy
- Precision and recall
- Squared Error
- Likelihood
- Posterior probability
- Cost/Utility
- Margin
- Entropy
- KL divergence

Examples of optimization

- Combinatorial optimization
  - e.g.: Greedy search
- Convex optimization
  - e.g.: Gradient descent
- Constrained optimization
  - e.g.: Linear programming

1.5 Supervised learning taxonomy

- Parametric vs Nonparametric
  - Parametric: fixed and finite number of parameters
  - Nonparametric: the number of parameters depends on thetraining set
- Frequentist vs Bayesian
  - Frequentist: use probabilities to model the sampling process
  - Bayesian: use probability to model uncertainty about the estimate

- Empirical Risk Minimization vs Structural Risk Minimization
  - Empirical Risk: Error over the training set
  - Structural Risk: Balance training error with model complexity
- Direct vs Generative vs Discriminative
  - Generative: learns the joint probability distribution $p(x, t)$
  - Discriminative: learns the conditional probability distribution $p(t|x)$

1.6 Learning approaches

Direct approach Learn directly f from D

Discriminative approach

- Model $p(t|x)$
- Marginalize to find $E[t|x] = \int t \cdot p(t|x)\, dt$

Generative approach

- Model $p(x, t)$
- Infer $p(t|x)$ (Bayes rule)
- Marginalize to find $E[t|x] = \int t \cdot p(t|x)\, dt$

2 Linear regression

Regression Learn an approximation of $f(x) : X \to \mathbb{R}$

- How to model f?
- How to optimize the approximation?
- How to evaluate the approximation?

Linear regression models f with linear functions

- easy to explain
- analytically solvable
- extendable to model non-linear relations
- base for more sophisticated models

First linear model

$$y(\vec{x}, \vec{w}) = \underbrace{w_0}_{\text{bias parameter}} + \sum_{j=1}^{D-1} w_j x_j = \vec{w}^T \cdot \underbrace{\vec{x}}_{(1, x_1, \ldots, x_{D-1})}$$

Sum of Squared Errors Error loss for linear regression

$$L(\vec{w}) = \frac{1}{2} \underbrace{\sum_{n=1}^{N} \left[y(x_n, \vec{w}) - t_n\right]^2}_{\text{Residual Sum of Squares}}$$

Residual Sum of Squares

$$RSS(\vec{w}) = \|\vec{\epsilon}\|_2^2 = \sum_{i=1}^{N} \epsilon_i^2$$

2.1 Linear models

We can define more complex models modeling non linearity: the regression model must be <u>linear in the parameters</u>, but the parameters can be non linear wrt the data.

Basis function is a function ϕ mapping data to parameters:

$$y(\vec{x}, \vec{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\vec{x}) = \vec{w}^T \cdot \underbrace{\vec{\phi}(\vec{x})}_{(1, \phi_1(\vec{x}), \ldots, \phi_{M-1}(\vec{x}))}$$

Some examples of basis functions:

.[Polynomial:] $\phi_j(x) = x^j$

Gaussian: $\phi_j(x) = e^{-\frac{(x-\mu_j)^2}{2\sigma^2}}$

Sigmoidal: $\phi_j(x) = \frac{1}{1+e^{\frac{\mu_j - x}{\sigma}}}$

Least Squares

$$L(\vec{w}) = \frac{1}{2} RSS(\vec{2}) = \frac{1}{2}(\vec{t} - \vec{\phi}\vec{w})^T(\vec{t} - \vec{\phi}\vec{w})$$

$$\frac{\partial L(\vec{w})}{\partial \vec{w}} = -\vec{\phi}^T(\vec{t} - \vec{\phi}\vec{w}) \qquad \frac{\partial^2 L(\vec{w})}{\partial \vec{w} \partial \vec{w}^T} = \vec{\phi}^T \vec{\phi}$$

$$\hat{\vec{w}}_{OLS} = \left(\vec{\phi}^T \vec{\phi}\right)^{-1} \vec{\phi}^T \vec{t}$$