Machine learning

Foundations

Corso di Laurea Magistrale in Informatica

Università di Roma Tor Vergata

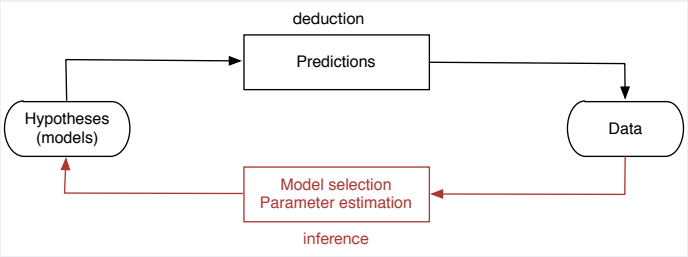
Prof. Giorgio Gambosi

a.a. 2021-2022

Machine learning: inductive approach

Learning of commonalities through analysis of a set of examples (training set), which is assumed to be available.

- A training set of n items is represented as a set of input vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, used to derive a model.
- If the purpose is item classification with respect to a collection of predefined classes, the training set also includes a target vector $\mathbf{t} = \{t_1, \ldots, t_n\}$, where the class of each training set item is specified.



Supervised learning

- We want to predict, given the values of a set (features) of an item $\mathbf{x}$, the unknown value of an additional feature target of the item
  - Target in $\mathbb{R}$: regression. Target in $\{1, \dots, K\}$: classification.
- General approach: defined (by means of learning from a set of examples) a model of the relation between feature and target values.
- The training set $\mathbf{X}, \mathbf{t}$ includes a feature vector $\mathbf{x}_i = \{x_{i1}, \dots, x_{im}\}$ and the corresponding target t_i for each item.
- The model could be:
  1. a function $y()$ which, for any item $\mathbf{x}$, returns a value $y(\mathbf{x})$ as an estimate of t
  2. a probability distribution which associates to each possible value $\overline{y}$ in the target domain, the corresponding probability $p(y = \overline{y}|\mathbf{x})$

Types of problems

Unsupervised learning

- We wish to extract, from a given collection of items dataset) $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, with no target associated, some synthetic information, such as:
  - subsets of similar items clustering)
  - the distribution of items in their domain (density estimation)
  - the projection, as informative as possible, of items on lower dimensional subspaces, that is, their characterization by means of a smaller set of features feature selection, feature extraction)
- A suitable model, of just the data features, is usually defined and applied also in the case of unsupervised learning.

Reinforcement learning

- We want to identify, in a given framework, a sequence of actions to be performed in order to maximize a certain profit
- As in supervised learning, no examples are given, but an environment is available which returns a profit in correspondance to the execution of any action

Machine learning framework: domains

Domain set $\mathcal{X}$: Set of objects we may wish to label. Each object is modeled as a vector of features. The number of features is the dimensionality of the problem

Label set $\mathcal{Y}$: Set of possible label values associated to objects in $\mathcal{X}$.

- $\mathcal{Y}$ continuous: regression
- $\mathcal{Y}$ discrete: classification

Training set $\mathscr{T}$: A set of object-label pairs: $\mathscr{T} = \{(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_n, t_n)\}$. We shall usually denote as $\mathbf{X}$ the matrix of objects (feature matrix), that is

$$\mathbf{X} = \begin{pmatrix} - & \mathbf{x}_1 & - \\ & \vdots & \\ - & \mathbf{x}_n & - \end{pmatrix}$$

and as $\mathbf{t}$ the vector of labels (target vector), that is

$$\mathbf{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_n \end{pmatrix}$$

Machine learning framework: learner output

Learner output: The learner (an algorithm A) is requested to return, for a given training set $\mathcal{T}$, a prediction rule (classifier, regressor) $A(\mathcal{T}) = h : \mathcal{X} \mapsto \mathcal{Y}$

Training objects generation model: We assume that the objects observed in the training set are sampled from $\mathscr{X}$ according to some probability distribution $\mathscr{D}_1$. That is, for any $\mathbf{x} \in \mathscr{X}$, $p_{\mathscr{D}_1}(\mathbf{x})$ is the probability that $\mathbf{x}$ is the next object sampled in the training set

Training targets generation model: In the general case, we assume the labels associated to the objects in the training set are generated according to a probability distribution $\mathscr{D}_2$ conditional on $\mathscr{X}$. That is, for any $t \in \mathscr{Y}$, $p_{\mathscr{D}_2}(t|\mathbf{x})$ is the probability that the observed label of object $\mathbf{x}$ in the training set is t. For the moment, we shall assume that the relation between object and label is deterministic, that is it exists an unknown function f such that $t_i = f(\mathbf{x}_i)$

Machine learning framework: prediction risk

Given any element $\mathbf{x} \in \mathcal{X}$:

Error: The error of a predictor h derives from the comparison of its prediction $h(\mathbf{x})$ and the correct target label y.

Loss: The comparison is performed by applying a predefined loss function $L : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$.

Risk of prediction: The error of a prediction $\hat{y}$ is defined in terms of prediction risk as given by applying the loss

$$\mathscr{R}(\hat{y}, \mathbf{x}) = L(\hat{y}, f(\mathbf{x}))$$

In the general case when only a probabilistic relation $p_{\mathscr{D}_2}(y|\mathbf{x})$ is assumed between label and target, this corresponds to

$$\mathscr{R}(\hat{y}, \mathbf{x}) = E_{\mathscr{D}_2}[L(\hat{y}, y)] = \int_{\mathcal{Y}} L(\hat{y}, y) \cdot p_{\mathscr{D}_2}(y|\mathbf{x}) dy$$

or, in the case of classification

$$\mathscr{R}(\hat{y}, \mathbf{x}) = E_{\mathscr{D}_2}[L(\hat{y}, y)] = \sum_{y \in \mathcal{Y}} L(\hat{y}, y) \cdot p_{\mathscr{D}_2}(y|\mathbf{x})$$

In this framework, the optimal prediction is the one which minimizes the risk,

$$y^*(\mathbf{x}) = \underset{\hat{y}}{\operatorname{argmin}}\, \mathscr{R}(\hat{y}, \mathbf{x}) = \underset{\hat{y}}{\operatorname{argmin}}\, E_{\mathscr{D}_2}[L(\hat{y}, y)]$$

that is,

$$y^*(\mathbf{x}) = \underset{\hat{y}}{\operatorname{argmin}}\, L(\hat{y}, f(\mathbf{x})) \qquad \text{in the simpler case}$$

$$y^*(\mathbf{x}) = \underset{\hat{y}}{\operatorname{argmin}}\, E_{\mathscr{D}_2}[L(\hat{y}, y)] = \underset{\hat{y}}{\operatorname{argmin}} \int_{\mathscr{Y}} L(\hat{y}, y) \cdot p_{\mathscr{D}_2}(y|\mathbf{x}) dy \qquad \text{in the general case}$$

in the general case, this is denoted as Bayes estimator.

However, observe that this approach cannot be applied since both the function f and the distribution $\mathscr{D}_2$ of $p(y|\mathbf{x})$ are assumed unknown.

The error of a predictor h is defined in terms of risk expected loss on all objects in $\mathcal{X}$

$$\mathcal{R}(h) = E_{\mathcal{D}_1, f}[L(h(\mathbf{x}), f(\mathbf{x}))] = \int_{\mathcal{X}} L(h(\mathbf{x}), f(\mathbf{x})) \cdot p_{\mathcal{D}_1}(\mathbf{x}) d\mathbf{x}$$

In the general case,

$$\mathcal{R}(h) = E_{\mathcal{D}_1, \mathcal{D}_2}[L(h(\mathbf{x}), y)] = \int_{\mathcal{X}} \int_{\mathcal{Y}} L(h(\mathbf{x}), y) \cdot p_{\mathcal{D}_1}(\mathbf{x}) \cdot p_{\mathcal{D}_2}(y|\mathbf{x}) d\mathbf{x} dy$$

Since $\mathscr{D}_1$ and $\mathscr{D}_2$ (or f) are not known, the risk can only be estimated from the data available (the training set $\mathscr{T}$).

Empirical risk: The risk can be estimated from the training set by estimating the expectation of the loss function as the average loss on the set.

$$\overline{\mathscr{R}}_{\mathscr{T}}(h) = \frac{1}{|\mathscr{T}|} \sum_{(x,t)\in\mathscr{T}} L(h(x), t)$$

Machine learning framework: from learning to optimization

The fundamental approach in machine learning is deriving a predictor h which (at least approximately) minimizes the empirical risk computed on the available training set.

A learning problem is then reduced to a minimization problem in some functional space $\mathscr{H}$, the set of all possible predictors h.

$$h^* = \operatorname*{argmin}_{h \in \mathscr{H}} \overline{\mathscr{R}}_{\mathscr{T}}(h)$$

Here, $\mathscr{H}$ is the set of hypotheses or inductive bias

Issues related to the inductive bias

The choice of the set of hypotheses is an important issue in ML:

- what is the effect of the structure and size of $\mathscr{H}$?
- how to define $\mathscr{H}$ in such a way to make it feasible to compute h^*?

Choice of $\mathscr{H}$

- The hypotheses class $\mathscr{H}$ can be viewed as reflecting some prior knowledge that the learner has about the task
  - a belief that one of the members of the class $\mathscr{H}$ is a low-error model for the task
- A trivial way of pursuing this goal would be to define a very rich class, that is assuming that many possible functions belong to $\mathscr{H}$
- As a limit, $\mathscr{H}$ could be defined just as the set of all functions $f : \mathcal{X} \mapsto \mathcal{Y}$

Problem with large $\mathscr{H}$:

- Assume a binary classification problem with training set $\mathscr{T} = (\mathbf{X}, \mathbf{t})$, with $0/1$ loss

$$L(y,t) = \begin{cases} 0 & \text{if } y = t \\ 1 & \text{otherwise} \end{cases}$$

  that is, the loss is 1 if the item is misclassified, 0 otherwise. As a consequence, the risk is the expected number of classification errors, while the empirical risk is the fraction of items in the training set which are misclassified.
- Assume $p(t = 1|\mathbf{x}) = \frac{1}{2}$ for $\mathbf{x} \in \mathscr{X}$, that is, the two classes have same size in the population

Consider the classification function defined as:

$$h(x) = \begin{cases} 1 & \text{if } \mathbf{x} = \mathbf{x}_i \in \mathbf{X}, t_i = 1 \\ 0 & \text{otherwise} \end{cases}$$

that is, h assigns to class 1 all items labeled as 1 in the training set. All other items are classified as 0.

Clearly, the empirical risk here is 0 by definition, but the risk is $\approx \frac{1}{2}$. When applied to a dataset randomly sampled from the population, the quality of h^* is the same of a function which randomly assigns items to classes.

This is called overfitting: the classification method behaves well on the training set, but poorly on new data from the population.

Choice of $\mathscr{H}$

With respect to $\mathscr{H}$, the following considerations can be done:

- If $\mathscr{H}$ is too large (complex), overfitting may occur: a function which behaves very well on the training set may be available which however performs poorly on new data
- If $\mathscr{H}$ is too small (simple), underfitting may occur: no function behaving in a satisfactory way, both on the training set and on new sets of data, is available in $\mathscr{H}$

This is related to the so-called bias variance tradeoff

Bias vs variance

The risk associated to the h^*, the predictor which minimizes the empirical risk, can be decomposed in two parts:

$$\mathscr{R}(h^*) = \epsilon_B + \epsilon_V$$

where:

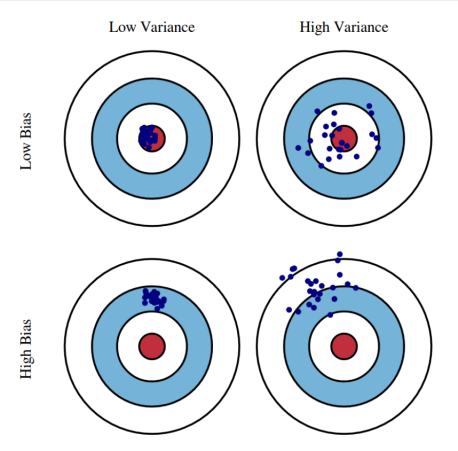
- ϵ_B is the minimum risk achievable by any $h \in \mathscr{H}$: this is only determined by the inductive bias, and independent from the training set. It is a property of the class of hypotheses considered with respect to the prediction task. This is called bias
- ϵ_V is the difference between the above minimum risk in $\mathscr{H}$ and the risk associated to the best predictor in $\mathscr{H}$ with respect to the training set: it is related to the fact that empirical risk minimization only provides an estimate of the best predictor achievable for the given inductive bias. It is a measure of how well the predictor computed from a particular training set approximates the best possible one. Its expectation with respect to all possible training sets is a measure of how much a predictor derived from a random training set may result in poorer performances with respect to the best possible one. This is called variance

Bias vs variance

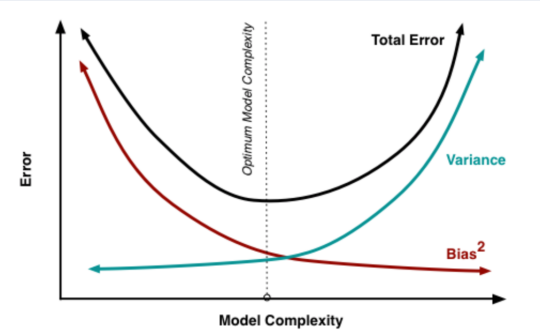
The choice of $\mathscr{H}$ is subject to a bias-variance tradeoff: higher bias tend to induce lower variance, and vice versa.

- High bias and low variance implies that all predictors which can be obtained from different training sets tend to behave similarly, with a similar risk (low variance). However, all of them then to behave poorly (high bias), since $\mathscr{H}$ is too poor to include a satisfactory predictor for the task considered. This results into underfitting
- Low bias and high variance implies that lot of predictors are available in $\mathscr{H}$, and among them a good one is usually avaliable (low bias). However, quite different predictors can be obtained from different training sets, which implies that it may easily happen that, while a very good performance can be obtained on the training set, the resulting predictor can behave quite differently and more poorly that the best possible one, which implies overfitting

Bias vs variance



Bias vs variance



Computing h^*

- The optimization required to derive h^* can be complex in the general case, when a function must be derived in a function space.
- Usually, the situation is made easier by considering $\mathcal{H}$ as a space of functions parameterized by a suitable set of coefficients (for example, all polynomials of degree at most d, for a given d): this results in a minimization to be performed over a set of d-dimensional points.
- That is, $\mathcal{H} = \{h_\theta | \theta \in \Theta\}$, where Θ is the coefficients domain and h is a function template, parameterized by elements in Θ
- Minimizing the Empirical risk results into computing

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \overline{\mathcal{R}}_{\mathcal{T}}(h_\theta)$$

Computing h^*

- In most cases, $\Theta = \mathbb{R}^d$ for some $d > 0$: in this case, the minimization of $\overline{\mathcal{R}}_{\mathcal{T}}(h_\theta)$ is unconstrained and a (at least local) minimum could be computed setting all partial derivatives to 0

$$\frac{\partial}{\partial \theta_i} \overline{\mathcal{R}}_{\mathcal{T}}(h_\theta) = 0$$

  that is, setting to zero the gradient of the empirical risk with respect to the vector of parameters θ

$$\nabla_\theta \overline{\mathcal{R}}_{\mathcal{T}}(h_\theta) = \mathbf{0}$$

- The analytical solution of this set of equations is usually quite hard
- Numerical methods can be applied

Gradient descent

- Gradient descent performs minimization of a function $J(\theta)$ through iterative updates of the current value of θ (starting from an initial value $\theta^{(0)}$) in the opposite direction to the one specified by the current value of the gradient $J'(\theta) = \nabla_\theta J(\theta)$

$$\theta^{(k+1)} = \theta^{(k)} - \eta \nabla_\theta J(\theta) \,|_{\theta=\theta^{(k)}}$$

  that is, for each parameter θ_i

$$\theta^{(k+1)} = \theta^{(k)} - \eta \frac{\partial J(\theta)}{\partial \theta_i}\Big|_{\theta=\theta^{(k)}}$$

- η is a tunable parameter, which controls the amount of update performed at each step

In Machine learning, minimization of the Empirical Risk is performed, hence gradient descent takes the form

$$\begin{aligned}\theta_i^{(k+1)} &= \theta_i^{(k)} - \eta \frac{\partial}{\partial \theta_i} \frac{1}{|\mathcal{T}|} \sum_{(x,t)\in\mathcal{T}} L(h_\theta(x), t)|_{\theta=\theta^{(k)}} \\ &= \theta_i^{(k)} - \frac{\eta}{|\mathcal{T}|} \sum_{(x,t)\in\mathcal{T}} \frac{\partial}{\partial \theta_i} L(h_\theta(x), t)|_{\theta=\theta^{(k)}}\end{aligned}$$

This is called batch gradient descent: observe that, at each step, all items in the training set must be considered

Batch gradient descent can be modified by performing the update, at each step, on the basis of the evaluation at a single item of the training set for single parameters,

$$\theta_i^{(k+1)} = \theta_i^{(k)} - \eta \frac{\partial}{\partial \theta_i} L(h_\theta(x_j), t_j)|_{\theta=\theta^{(k)}}$$

Mini-batch gradient descent

An intermediate case is the one when a subset of the items in the training is considered at each step

$$\theta_i^{(k+1)} = \theta_i^{(k)} - \frac{\eta}{|B_r|} \sum_{(x,t)\in B_r} \frac{\partial}{\partial \theta_i} L(h_\theta(x), t)|_{\theta=\theta^{(k)}}$$

This is called mini-batch gradient descent

Momentum gradient descent

- Based on a physical interpretation of the optimization process: a body of mass $m = 1$ is moving on the surface of a cost function $J(\theta)$, with potential energy $U(\theta) = \eta J(\theta)$ and weight force (or acceleration, since $m = 1$) $F(\theta) = -\nabla U(\theta) = -\eta \nabla J(\theta)$, at any point θ
- In gradient descent, the movement of the body is determined by the acceleration at that point, that is by the gradient $\nabla J(\theta)$
- In momentum gradient descent, the velocity $v(\theta)$ of the body is considered: the movement of the body is determined by the velocity, that is,

$$\theta^{(k+1)} = \theta^{(k)} + v^{(k+1)}$$

with the velocity changing as determined by the acceleration

$$v^{(k+1)} = v^{(k)} - \eta \nabla J(\theta^{(k)})$$

This results into

$$v^{(k+1)} = -\eta \nabla J(\theta^{(k)}) + v^{(k)} = -\eta \nabla J(\theta^{(k)}) - \eta \nabla J(\theta^{(k-1)}) + v^{(k-1)} = \cdots$$
$$= -\eta \sum_{i=0}^{k} \nabla J(\theta^{(i)}) + v^{(0)}$$

$$\theta^{(k+1)} = \theta^{(k)} + v^{(k+1)} = \theta^{(k)} - \eta \sum_{i=0}^{k} \nabla J(\theta^{(i)}) + v^{(0)}$$

Nesterov gradient descent

- The same approach of momentum gradient descent is applied, with the gradient estimation performed not at the current point $\theta^{(k)}$, but approximately at the next point $\theta^{(k+1)}$
- The approximation derives by considering $\theta^{(k)} + \gamma v^{(k)}$ instead of $\theta^{(k+1)}$
- The updates of v and θ are considered in advance with respect to momentum GD

$$v^{(k+1)} = \gamma v^{(k)} + \eta \nabla J(\tilde{\theta}^{(k)}) = \gamma v^{(k)} + \eta \nabla J(\theta^{(k)} + \gamma v^{(k)})$$
$$\theta^{(k+1)} = \theta^{(k)} + v^{(k+1)}$$

Adagrad

- In Adagrad, different learning rates are applied to the different parameters θ_i: larger gradients in the preceding steps results into smaller rates
- in particular, the learning rate of θ_j at step k is defined as

$$\eta_j^{(k)} = \frac{\eta}{\sqrt{G_{j,k} + \varepsilon}}$$

  where η is a constant, $G_{j,k} = \sum_{i=0}^{k} g_{j,i}^2$ is the sum of the squared past gradients $g_{j,i} = \frac{\partial J(\theta, X)}{\partial \theta_j}\Big|_{\theta=\theta^{(i)}}$ and ε is a small *smoothing* constant, to deal with possible null denominators
- The update of θ_j at step $k+1$ is then

$$\theta_j^{(k+1)} = \theta_j^{(k)} - \frac{\eta}{\sqrt{G_{j,k} + \varepsilon}} g_{j,k}$$

- Learning rates decrease at each step, with the ones associated to parameters which had large gradients in the past decreasing more: that is, parameters which tended to large variation at each step are more "pushed" towards a stable value

Adadelta

- In Adadelta, past gradients are considered with a decreasing relevance of long past ones:
- A (decay) is applied my means of a coefficient $0 < \gamma < 1$

$$G_{j,k} = \gamma G_{j,k-1} + (1 - \gamma) g_{j,k}^2$$

Second order methods

- Maxima (or minima) of $J(\theta)$can be found by searching points where the gradient (all partial derivatives) zero
- Any iterative method to compute zeros of a function (such as Newton-Raphson) can the be applied on the gradient $\nabla J(\theta)$
- Such methods rely on the gradient of the function considered, hence, the gradient of $\nabla J(\theta)$ (that is, the Hessian) must be computed

$$H_{ij}(J(\theta)) = \frac{\partial^2 J(\theta)}{\partial \theta_i \partial \theta_j}$$

- At each step, the following iteration is applied (in case Newton-Raphson method is used)

$$\theta^{(k+1)} = \theta^{(k)} - \left(H(J(\theta))^{-1} \nabla J(\theta)\right)\Big|_{\theta_{(k)}}$$

Probabilistic approaches

As done before, we assume that the observed dataset (features and target) has been derived by randomly sampling:

- $\mathscr{X}$ according to the probability distribution $p_{\mathscr{D}_1}(x)$ (usually the uniform distribution)
- $\mathscr{Y}$ according to the conditional distribution $p_{\mathscr{D}_2}(y|x)$

1. we may then consider a class of possible conditional distributions $\mathscr{P}$ and
2. select (infer) the "best" conditional distribution $p^* \in \mathscr{P}$ from the available knowledge (that is, the dataset), according to some measure q
3. given any new item x, apply $p^*(y|\mathbf{x})$ to assign probabilities for each possible value of the corresponding target
4. an independent decision strategy must be applied to $p^*(y|\mathbf{x})$ to return a specific prediction $h(\mathbf{x})$

Inferring a best distribution

- how to define the class of possible conditional distributions $p(y|\mathbf{x})$?
  - usually, parametric approach: distributions defined by a common (arbitrary) structure and a set of parameters
- what is a measure $q(p, \mathcal{T})$ of the quality of the distribution (given the dataset $\mathcal{T} = (\mathbf{X}, \mathbf{t})$)?
  - this is related to how a dataset generated by randomly sampling from $\mathcal{D}_1$ (usually uniform) and $\mathcal{D}_2$ could be similar to the available dataset $\mathcal{T}$

A different approach

Instead of finding a best distribution $p^* \in \mathscr{P}$ and use it to predict target probabilities as $p^*(y|\mathbf{x})$ for any element $\mathbf{x}$, we could

- consider for each possible conditional distribution $p \in \mathscr{P}$ its quality $q(p, \mathscr{T})$
- compose all conditional distributions $p(y|\mathbf{x})$ each weighted by its quality $q(p, \mathscr{T})$ (for example by means of a weighted averaging)
- apply the resulting distribution

Different strategies

Assume q takes the form of a probability distribution (of probability distribution)

- first approach: take the modal value (the distribution of maximum quality) and apply it to perform predictions
- second approach: compute the expectation of the distributions, wrt the probability distribution q