Theoretical Machine Learning

With applications to finance

Dhruv Madeka

Quantitative Researcher
Bloomberg LP

October 18, 2016

Outline

Outline

What is Machine Learning?

Definition

Machine Learning can be broadly defined as computational methods that allow for automatic improvement of a task (learning) through experience.

What is it used for?

A better question might be what isnt it used for?

- Classification
- Regression
- Clustering
- Ranking
- Manifold Learning

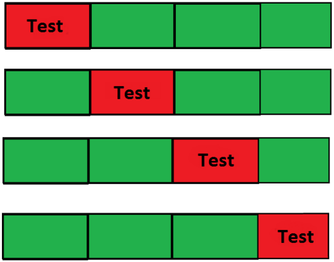
Commonly used terms

- Examples
- Features
- Labels
- Training Sample
- Validation Sample
- Test Sample
- Loss Function
- Hypothesis Set
- Cross-Validation

Cross-Validation

For a sample size of m, an n-fold cross validation $\implies$ sample size of $m - \frac{m}{n}$

Figure: Cross-Validation

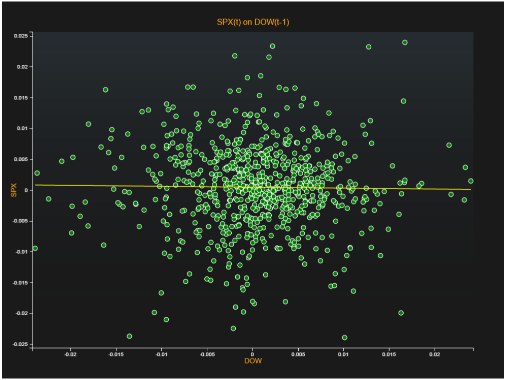


- Large n $\implies$ low bias-high variance
- Small n $\implies$ high bias-low variance

Supervised Learning

The learner receives a set of labeled training data and makes predictions for all unseen points.

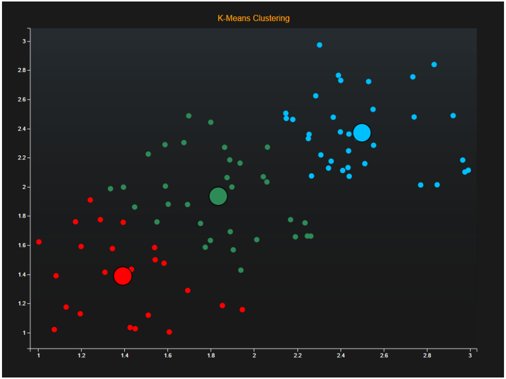
Figure: Regression of SPX on lagged DOW



Unsupervised Learning

The learner receives a set of unlabeled training data and tries to infer the density (or properties) of the points

Figure: K-Means as an example of unsupervised learning

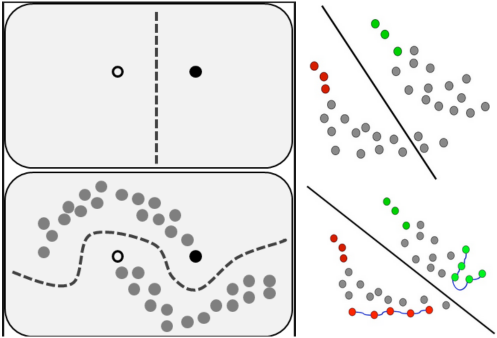


Supervised vs Unsupervised Learning

- Supervised learning attempts to learn through guidance (errors of each observation) the conditional density of a variable Y given another variable X
- Unsupervised learning tries to infer properties of the underlying joint density of a random vector X without the help of a teacher (degree of error for each observation)

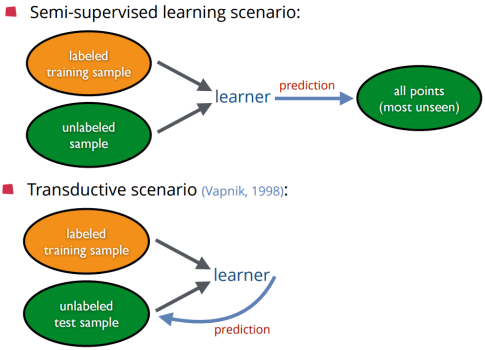
Semi-supervised Learning

The learner receives both labeled and unlabeled points and has to make inferences about all unseen points.



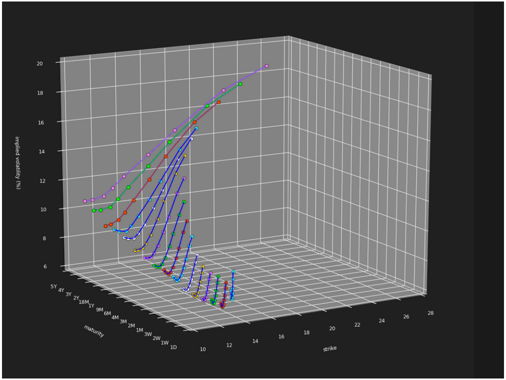
Transduction

The learner receives both labeled and unlabeled training points and has to make inferences only on the test points.



Transduction

The interesting cases for transduction are when the learner can perform better than a supervised or unsupervised learner with the labeled and unlabled points respectively.



On-line Learning

Rather than making distributional assumptions on the data, on-line learning measures performance by using a mistake model or the notion of regret.

- At step t, receive instance x_t (or instance and expert advice $\bar{y}_{t,i}$ $\forall i \in [1, N]$
- Predict a label $\hat{y}_t$
- Receive a label $y_t \in Y$
- Incur loss $L(\hat{y}_t, y_t)$

Objective: Minimize regret or cumulative loss:

$$\sum_{t=1}^{T} L(\hat{y}_t, y_t) - \min_i \sum_{t=1}^{T} L(\bar{y}_{t,i}, y_t) \tag{1}$$

Reinforcement Learning

- Reinforcement Learning refers to the scenarios where the learner collects information through a course of actions by interacting with the environment.
- The learner receives two things in response to the action: his current state, and a real value reward which needs to be maximized.

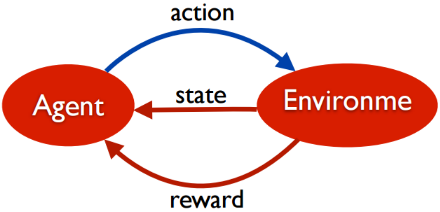


Figure: Model of Reinforcement Learning

What is Deep Learning?

Linear models

Transformed linear models

Non-linear models

Deep Learning

???

- > 1 non-linear transformations applied to data
- sometimes, Deep Learning $=$ Neural Networks
- sometimes, Deep Learning $=$ Probabilistic Graphical Models

Active Learning

- Typically, a learner receives an entire labeled sample $((x_1, y_1), ..., (x_N, y_N))$
- An active learner receives a sample $(x_1, ..., x_N)$ and can request each label
- One objective might be to request fewer labels than a passive learner (useful when labels are expensive)

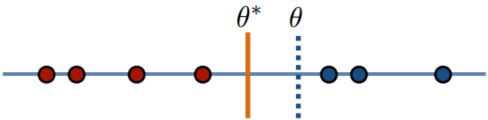


Figure: Favorable example: Binary classification in $\mathbb{R}$

Outline

Definitions

- $\mathcal{X}$: The input space is the set of all possible examples or instances
- $\mathcal{Y}$: The set of all possible labels or target values (target space)
- c: $\mathcal{X} \rightarrow \mathcal{Y}$ is a mapping from $\mathcal{X}$ to $\mathcal{Y}$
- C: Set of concepts we may wish to learn
- D: Example distribution (example are assumed to be i.i.d.)

Definitions

- H: The learner considers a fixed set of possible concepts H called the hypothesis space
- S: The learner receives a fixed sample assumed to be drawn from D (and assumed i.i.d.) and labels c(S)
- $\mathcal{A}$: The algorithm of the learner is a mapping: $S \to h_S \in H$

Consistency Model

Consistency Model

We say that a concept class C is learnable under the consistency model if $\exists$ an algorithm $\mathcal{A}$, which when given any set of labeled examples, finds a concept c that is consistent with all the examples, or says correctly that one does not exist.

Examples

Example

The set of monotone conjuctions can be learned through a bit-wise conjuction algorithm

Example

How does one learn the set of monotone disjunctions?

Example

The set of axis-aligned rectangles can be learned by selecting the tightest rectangle that fits the sample

Issues

- The notion of learnability is entirely sample dependent. It says nothing about the accuracy of the model on new data.
- Concept classes can be learnable under the consistency model but certain subsets of these classes might not be (e.g. 2-DNF)

Error Definitions

Generalization Error

Given a hypothesis $h \in H$, a target concept $c \in C$, and an underlying distribution D, the generalization error or risk of h is defined by:

$$R(h) = \mathop{\mathbb{E}}_{x \sim D}[L(h(x), c(x))] \tag{2}$$

Here L denotes a loss function, $L : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}_+$

Empirical Error

Given a hypothesis $h \in H$, a target concept $c \in C$, and a sample $S = ((x_1, y_1), ..., (x_N, y_N))$, the empirical error or risk of h is defined by:

$$\hat{R}(h) = \frac{1}{N} \sum_{i=1}^{N} L(h(x_i), y_i) \tag{3}$$

PAC Learning

PAC-learning

A concept class C is said to be PAC-learnable if $\exists \mathcal{A}$ and a polynomial function poly(., ., ., .) such that for any $\epsilon > 0, \delta > 0$, $\forall D$ on $\mathcal{X}$ and for any target concept $c \in C$, the following holds for any sample size $m \geq \text{poly}(\frac{1}{\epsilon},\ \frac{1}{\delta},\ n,\ \text{size}(c))$:

$$\Pr_{S \sim D^m}[R(h_s) \leq \epsilon] \geq 1 - \delta \tag{4}$$

If $\mathcal{A}$ further runs in $\text{poly}(\frac{1}{\epsilon},\ \frac{1}{\delta},\ n,\ \text{size}(c))$, then C is said to be efficiently PAC-learnable.

Example

Example

Learning on a line Assume $\mathcal{X} = \mathbb{R}$ and $C =$ positive half lines. Find $c : [c, \infty)$ are labeled $+$.

Learning bounds - consistent case

Theorem

Let H be a finite set of hypothesis functions mapping $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{A}$ be an algorithm that for any target concept $c \in H$ and i.i.d sample S returns a consistent hypothesis h_S $(\hat{R}(h_S) = 0)$. Then, for any $\epsilon,\ \delta > 0$, the inequality $\underset{S \sim D}{Pr}[R(h_s) \leq \epsilon] \geq 1 - \delta$ holds if:

$$m \geq \frac{1}{\epsilon}\left(log|H| + log\frac{1}{\delta}\right) \tag{5}$$

Equivalently, for any $\epsilon,\ \delta > 0$, with probability $\geq 1 - \delta$:

$$R(h_S) \leq \frac{1}{m}\left(log|H| + log\frac{1}{\delta}\right) \tag{6}$$

Concentration Inequalities

Theorem

Hoeffding's Inequality

Let $X_1, ..., X_m$ be independent random variables with X_i taking values in $[a_i, b_i]$ $\forall i \in [1, m]$. Then for any $\epsilon > 0$, the following inequalities hold for $S_m = \sum_{i=1}^{M} X_i$.

$$Pr[S_m - \mathbb{E}[S_m] \geq \epsilon] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}}$$

$$Pr[S_m - \mathbb{E}[S_m] \leq -\epsilon] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}}$$

Typically used to analyze generalization error bounds.

McDiarmid's Inequality

Theorem

Let $(X_1, ..., X_m) \in \mathcal{X}^m$ *be a set of* $m \geq 1$ *independent random variables and assume* $\exists (c_1, ..., c_m) > 0$ *such that* $f : \mathcal{X}^m \to \mathbb{R}$ *satisfies the following conditions:*

$$|f(x_1, ..., x_i, ..., x_m) - f(x_1, ..., x_i^{'}, ..., x_m)| \leq c_i$$

$\forall i \in [1, m]$ *and any points* $x_1, ..., x_m, x_i^{'} \in \mathcal{X}$. *Let* $f(S)$ *denote* $f(X_1, ..., X_m)$, *then,* $\forall \epsilon > 0$, *the following inequalities hold:*

$$Pr[f(S) - \mathbb{E}[f(S)] \geq \epsilon] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{m}(c_i)^2}}$$

$$Pr[f(S) - \mathbb{E}[f(S)] \leq -\epsilon] \leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{m}(c_i)^2}}$$

Typically used to analyze bounds on Rademacher complexity.

Learning bounds - inconsistent case

Theorem

Let H be a finite set of hypothesis functions mapping $\mathcal{X}$ to $\mathcal{Y}$. Then, for any $\epsilon,\ \delta > 0$, the inequality:

$$\forall h \in \mathcal{H},\ R(h) \leq \hat{R}(h) + \sqrt{\frac{log|\mathcal{H}| + log\frac{2}{\delta}}{2m}} \tag{7}$$

Or:

$$\mathbb{P}\Big[\exists h \in H : |\hat{R}(h) - R(h)| > \epsilon\Big] \leq 2|\mathcal{H}|e^{-2m\epsilon^2} \tag{8}$$

The control of the empirical error versus the size of the hypothesis space, is another statement of *Occam's Razor*, which says that plurality should not be posited without necessity.

Stochastic Scenarios

What if both x and y are random variables?

$$R(h) = \Pr_{(x,y)\sim D}[h(x) \neq y] = \mathop{\mathbb{E}}_{(x,y)\sim D}[1_{h(x)\neq y}]$$

PAC Learning $\rightarrow$ Agnostic PAC learning.
Replace (4) by:

$$\Pr_{S\sim D^m}[R(h_s) - \min_{h\in\mathcal{H}} R(h) \leq \epsilon] \geq 1 - \delta \tag{9}$$

Infinite Hypothesis Spaces

- Obviously the bounds above are totally uninformative for infinite hypothesis spaces
- Clearly (case of the separating point) learning is possible even with infinite hypothesis spaces
- Useful measures of complexity for infinite hypothesis spaces include VC dimension and Rademacher complexity

Growth Function

Dichotomies and Shattering

Given a hypothesis $h \in \mathcal{H}$, a dichotomy is one possible way of labeling a sample S using a hypothesis in $\mathcal{H}$.
A sample S is said to be shattered by a hypothesis set $\mathcal{H}$ when $\mathcal{H}$ realizes all possible dichotomies of S.

So:

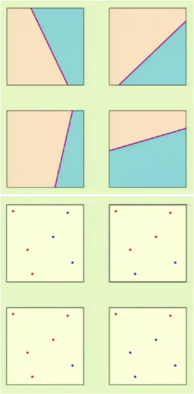
$$\text{Hypothesis: } h : \mathcal{X} \to \mathcal{Y}$$
$$\text{Dichotomy: } h_S : \{x_1, ..., x_N\} \to \mathcal{Y}$$

Break-Point

If no set larger than size k can be shattered by a hypothesis set $\mathcal{H}$, then k is said to be the break point of $\mathcal{H}$. For example, positive half rays (k=2), 2D perceptrons (k=4).

Intuition

Typically, we have an input space which is over a continuum of points. However, if we put an opaque sheet on top of it, and put holes in the sheet at the sample points, we no longer have this continuum.



Growth Function

Growth Function

For an unlabeled sample S $= \langle x_1, ..., x_N \rangle$, define a behaviour set $\Pi_{\mathcal{H}}(\mathcal{S}) = \{\langle h(x_1), ..., h(x_N) \rangle : h \in \mathcal{H}\}$ and a growth function $\Pi_{\mathcal{H}}(m) = \max_{|\mathcal{S}|=m} |\Pi_{\mathcal{H}}(\mathcal{S})|$

Intuition

We can see that typically

$$\mathbb{P}[R(h) - \hat{R}(h) > \epsilon] \leq 2|\mathcal{H}|e^{-2\epsilon^2 N} \tag{10}$$

We seek to replace this by:

$$\mathbb{P}[R(h) - \hat{R}(h) > \epsilon] \leq 4\Pi_{\mathcal{H}}(2N)e^{-\frac{1}{8}\epsilon^2 N} \tag{11}$$

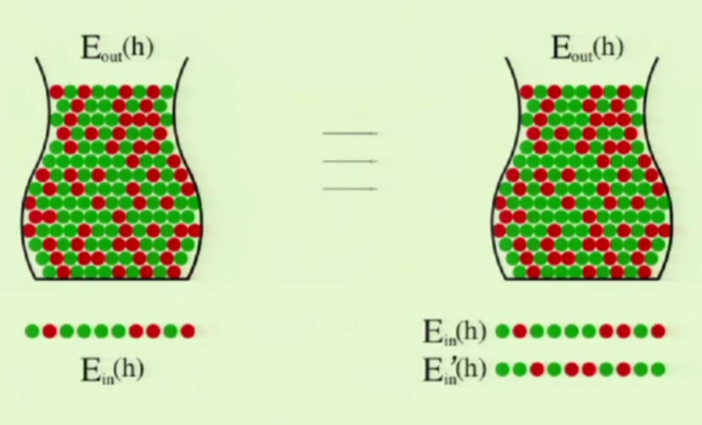
We can do this if $\Pi_{\mathcal{H}}(N)$ is polynomial in N. And actually:

$$\Pi_{\mathcal{H}}(N) \leq \sum_{i=1}^{k-1} \binom{N}{i} \tag{12}$$

where k is the break point of the hypothesis set.

Why did the coefficients change?

The only subtlety in the argument from dichotomies comes because $R(h)$ is not sample dependent, but rather distribution dependent. So, the proof looks at it in this way:



Bounds on the error through the growth function

The worst case is that we need a hypothesis set with size equal to all possible functions on m points, which is 2^m. So there are only two possible cases for $\Pi_{\mathcal{H}}(m)$. Either $\Pi_{\mathcal{H}}(m) = 2^m$ (learning is hard) or $\Pi_{\mathcal{H}}(m) = O(m^d)$ where d is the Vapnik Chervonenkis (VC) dimension of $\mathcal{H}$.

Theorem

With probability $\geq 1 - \delta$. $\forall h \in \mathcal{H}$, if h is consistent, then:

$$err_D(h) \leq O\left(\frac{ln(\Pi_{\mathcal{H}}(2m)) + ln\frac{1}{\delta}}{m}\right) \tag{13}$$

VC Dimension

VC Dimension

The cardinality of the largest set that can be shattered by a hypothesis set $\mathcal{H}$ is termed the Vapnik Chervonenkis (VC) dimension of the hypothesis set.

VC Dimension

- Independent of the input distribution
- Independent of the learning algorithm
- Independent of the target function

Examples

- Positive rays: d=1
- 2D Perceptrons: d=3
- Perceptrons: d=dimension+1

Typically, proofs for the VC dimension (say d) require that you prove any set of d can be shattered, and that no set of d+1 can be shattered. E.g. Radon's Theorem helps us show that any set X of $d+2$ points $\in \mathbb{R}^d$ cannot be shattered by a perceptron.

Radon's Theorem

Theorem

Any set X of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two subsets X_1 and X_2 such that the convex hulls of X_1 and X_2 intersect.

Observe that when two sets can be partitioned by a hyperplane, so can their convex hulls. Thus, if the two convex hulls intersect, X_1 and X_2 cannot be separated by a hyperplane and X is not shattered.

An (Admittedly Trivial) extension to Sauer's Lemma

We know from earlier that:

$$
\begin{aligned}
\Pi_{\mathcal{H}}(N) &\leq \sum_{i=0}^{d} \binom{N}{i} \\
&\leq \sum_{i=0}^{d} \binom{N}{i} \left(\frac{N}{d}\right)^{d-i} \\
&\leq \sum_{i=0}^{N} \binom{N}{i} \left(\frac{N}{d}\right)^{d-i} \\
&= \left(\frac{N}{d}\right)^{d} \sum_{i=0}^{N} \binom{N}{i} \left(\frac{d}{N}\right)^{i} \\
&= \left(\frac{N}{d}\right)^{d} \left(1 + \frac{d}{N}\right)^{N} \leq \left(\frac{N}{d}\right)^{d} e^{d}
\end{aligned}
$$

VC Dimension Generalization Bounds

We know from the growth function bounds that:

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{2 \log \Pi_{\mathcal{H}}(m)}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{14}$$

Using this in combination with Sauer's lemma:

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{15}$$

Rademacher Complexity

Empirical Rademacher Complexity

Given a training sample $S = (x_1, ..., x_N)$, a hypothesis set H, the empirical Rademacher complexity of H is defined by:

$$\bar{R}_N(H) = \mathbb{E}_\sigma \left[\max_{h \in H} \frac{2}{N} \sum_{i=1}^{N} \sigma_i h(x_i) \right] \tag{16}$$

where $\sigma_i = (\sigma_1, ..., \sigma_N)$, $\mathbb{P}(\sigma_i = +1) = 0.5$ and $\mathbb{P}(\sigma_i = -1) = 0.5$. $h : \mathcal{X} \to [0, 1]$.

The Rademacher Complexity of a hypothesis set is then defined as $\bar{R}(H) = \mathbb{E}_S[\bar{R}_N(H)]$

Rademacher Complexity Bounds

We can bound the generalization error in the following way, $\forall\delta > 0$, with probability at least $1-\delta, \forall h \in H$:

$$R(h) \leq \hat{R}_S(h) + \bar{R}_N(H) + C\sqrt{\frac{1}{N} \ln \frac{2}{\delta}} \tag{17}$$

$$R(h) \leq \hat{R}_S(h) + \frac{\bar{R}_N(G)}{2} + C\sqrt{\frac{1}{N} \ln \frac{2}{\delta}} \tag{18}$$

(The second equation represents a bound for classification problems.) Finally, for finite and infinite hypothesis sets, the bounds are as follows:

$$\bar{R}_N(\mathcal{F}) \leq 2N\sqrt{\frac{2 \log |\mathcal{F}|}{N}} \tag{19}$$

$$\bar{R}_N(\mathcal{F}) \leq 2N\sqrt{\frac{2d}{N} \log\left(\frac{2eN}{d}\right)} \tag{20}$$

For Further Reading I

[3] [2] [1]

Machine learning course.
http://work.caltech.edu/telecourse.html.

Sanjoy Dasgupta.
The two faces of active learning.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar.
Foundations of machine learning.
2012.