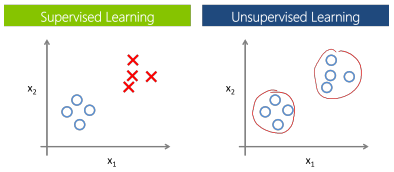
ML/Stats Cheatsheet

Compiled by Daniel Fernandez (used template from http://wzchen.com and Joe Blitzstein).

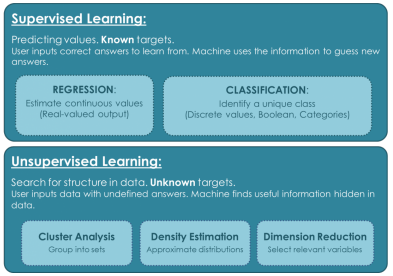
Last Updated December 29, 2016

Machine Learning Definitions

Supervised and Unsupervised Learning



Let us first define our response as a vector of dimension n, Y_n; and our predictors/features/variables as a matrix of dimension $n \times p$, $X_{n\times p}$ (i.e., p predictors).
In Supervised Learning one aim to train a model to classify/predict a categorical/numerical output Y, using features/variables X.
In Unsupervised Learning one aims to find patterns/summaries/groups in the given data - common tasks are clustering, and dimensionality reduction.

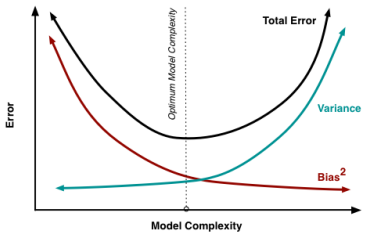


Fundamental Ideas

Curse of Dimensionality as the number of features grows one need exponentially more data points to provide local generalizations (i.e., KNN).

Bias and Variance Tradeoff The mean squared error (loss) of my model, $\hat{y}$ fit to the real data, y, exhibits a tradeoff between overparametrized models (e.g., splines) exhibiting low-bias, high-variance; and underparametrized models exhibiting high bias, low variance. Mathematically,

$$E\left[(\hat{y}-y)^2\right] = \text{Var}(\hat{y}) + [Bias(\hat{y})]^2$$



Train, validation and test In order to prevent overfitting, and having better measures of the model performance OOS one technique is to divide the set of data (X, Y) into train, validate, test. The train-validate set can be done using K-fold cross validation, or rolling cross-validation (in case of time series) along with hyper-parameter optimization (grid, bayesian or random search). One can optimize the model hyperparameters using The test set ensures that we have an unbias and objective measure of the error after the model hyperparameters have been optimized.

Large p, small n Problem . When $p >> n$ it is commonly called an large p, small n problem. I.e., a large feature space for a relatively small number of observations. Methods to tackle this problem rely on variable selection, and dimension reduction techniques. Examples easily arise in statistical genetics (large number of genes/SNPs ($20k/600k$ in humans), small number of observations of people with a given disease.

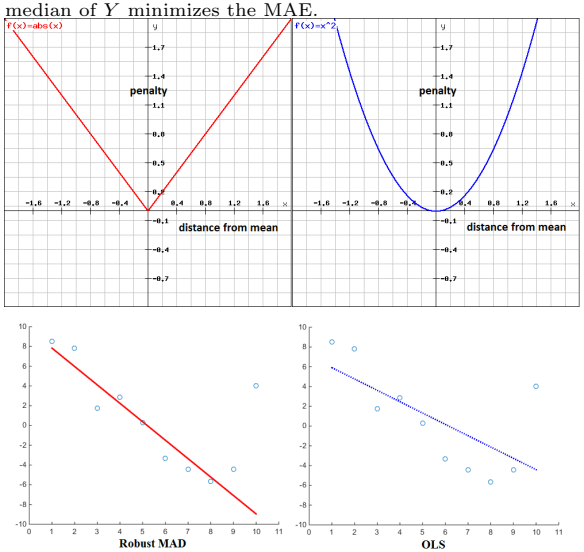
Admisibility . An admissible decision rule is a rule for making a decision such that there is not any other rule that is always "better" (as defined by MSE for example) than it.

Performance Metrics

continuous response

Mean Squared Error (MSE) $E\left[(y-\hat{y})^2\right]$.
It weights larger differences (outliers) more heavily (squares are larger as the error gets larger).
The mean of Y (when no X) minimizes the MSE.

Mean Absolute Error (MAE) . $E\left|y-\hat{y}\right|$.
It is less sensitive to large differences.
The median of Y minimizes the MAE.



Other measures are the Mean Absolute percentage error (MAPE), or the mean squared percentage error (MSPE).
When taking the median the sensitivity to outliers is further reduced, but it comes with an intuitive caveat. Remember that the median of the values 1, 2 and 3, and the median the values 1, 2 and 999, are both 2. If the numbers in the two series above were the amount of money you would risk to lose as a consequence of participating in two different investment alternatives over the next day, evaluating the risk of these two investments by median loss would clearly be irrational.

categorical response

binary response

Information Theory

Information Content The information content is related to the number of binary decisions required to find the information. I.e., Is it head or tail? $\log_2(2) = 1$. Which nucleotide? $\log_2(4) = 2$.
Generally, for a discrete random variable taking value i, $I(p_i) = -\log_2(p_i)$.
properties.
(1) $I(p)$ is anti-monotonic
(2) $I(0)$ is undefined,
(3) $I(p) \geq 0$ - information is a non-negative quantity
(4) $I(1) = 0$ - events that always occur contain no information
(5) $I(p_1 \times p_2) = I(p_1) + I(p_2)$ - information due to indenpendent events is additive.

Entropy Entropy quantifies the amount of uncertainty inherent in the value of a random variable (or the outcome of a random process). Specifically, it is the average information of all possible outcomes of the random variable.

$$H(X) = E[I(X)] = E[-\log p(x)] = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$$

Joint Entropy The entropy of the pairing (X, Y). If they are independent then it is the sum of $H(X) + H(Y)$.

$$H(X,Y) = E[I(X,Y)] = E_{X,Y}[-\log p(x,y)] = -\sum_{x,y} p(x,y)\log_2 p(x,y)$$

Conditional Entropy Or, conditional uncertainty of Y given X.

$$\begin{aligned} H(Y|X) &= E_X[H(Y|x)] = -\sum_x p(x)H(Y|x) \\ &= -\sum_x p(x)\sum_y p(y|x)\log p(y|x) \\ &= -\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)} \end{aligned}$$

$$H(Y|X) = H(X,Y) - H(X)$$

Mutual Information It measures the amount of information that can be obtained about one random variable by observing another random variable.

$$I(X,Y) = H(Y) - H(Y|X)$$

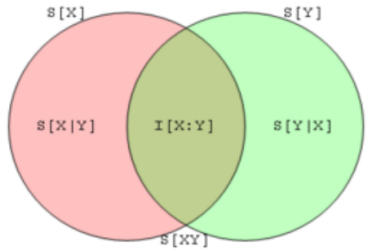
It is symmetric in X and Y,

$$I(X,Y) = I(Y,X) = H(X) + H(Y) - H(X,Y)$$

It can be expressed as the average KL (information gain) between the posterior probability of X given the value of Y, and the prior distribution on X.

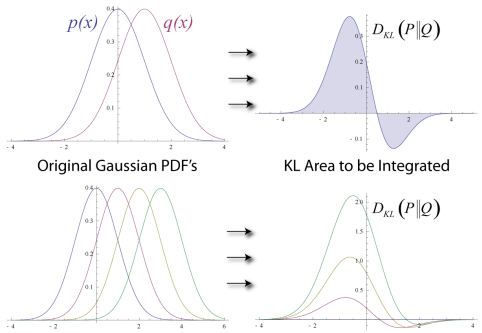
$$I(X,Y) = E_{p(y)}[D_{KL}(p(X|Y=y)||p(X))] = D_{KL}(p(X,Y)||p(X)p(Y))$$

Graphically,



Kullback-Leibler divergence, or information gain, or relative entropy.
I must transmit Y. How many bits on average would it save me if both ends of the line knew X?.
It is also interpreted as the difference between two probability distributions P and Q. It is not symmetric in P and Q. In applications, P typically represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution, while Q typically represents a theory, model, description, or approximation of P.
The KullbackLeibler divergence is sometimes also called the information gain achieved if P is used instead of Q. In Bayesian statistics the KullbackLeibler divergence can be used as a measure of the information gain in moving from a prior distribution to a posterior distribution.
It is also called the relative entropy of P with respect to Q.

$$D_{KL}(p(X)||q(X)) = \sum_x -p(x)\log q(x) - \sum_x -p(x)\log p(x)$$
$$= \sum_x p(x)\log\frac{p(x)}{q(x)}$$

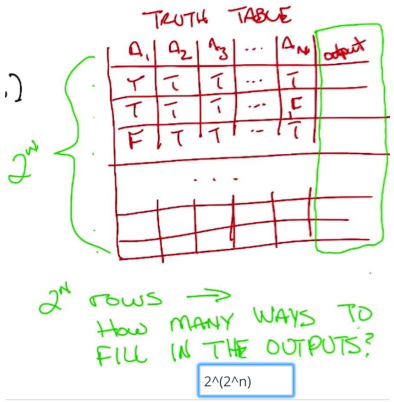


The assymetry is clear from the picture. We integrate w/r to p(x)

Linear Regression

Decision Trees

Tree-based methods partition the feature space into a set of rectangles and then fit a simple model in each one.
For binary X and binary Y, one would have $n-$OR, any, (linear complexity), and $n-$XOR, odd parity, (2^p exponential complexity).
Thus, the space of all possible deicsion trees is extremely large, encompassing 2^{2^n} possible trees (double exponential, very fast grow).



One cannot guarantee to find the optimal tree, best one can do is to use greedy-search methods. The most popular methods are CART and $ID3 \to C4.5 \to C5.0$.

Algorithm 1: ID3 Top Down Learning Algorithm

```
while examples not perfectly classified do
    Pick X_j that "best" separates Y
    Assign X_j as decision feature for node
    For each value of X_j create a descendant of node
    Assign training examples Y to each leaf descendant of node
end
```

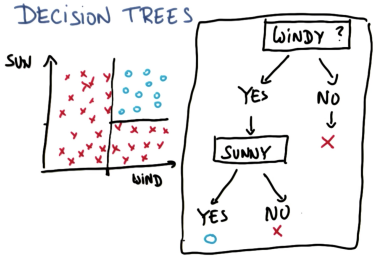
Maximum Gain It is defined as the gain in separation by adding a new node X_j to the decision tree.
Common measures are the gini index, and the information gain.
The information gain is defined as,

$$IG = H(parent) - [weighted_average]H(children)$$

This measure has a major selection bias for X_j categorical that takes multiple values.

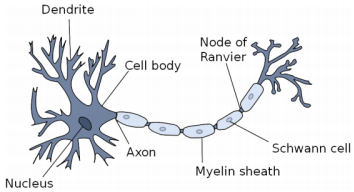
Stop criterion and overfitting Check accuracy in the cross validation set before decide to expand the tree. Alternatively, do the full tree, and then do prunning.

Overall Considerations Decision trees exhibit restriction bias, i.e., one only consider functional forms in the form of a tree.
They also exhibit a preference/inductive bias (certain trees are prefered more than others): Trees with good splits at the top, prefers ones that give better classification accuracy.
For continuous features, one can take ranges, and use them more than one time in the tree.
For regression (instead of classification) the split criteria based on MSE, and the output could be a local linear fit.

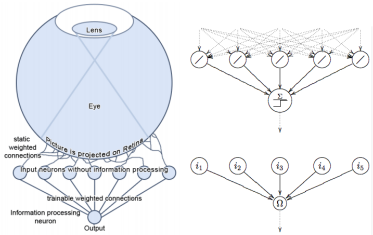


Neural Networks

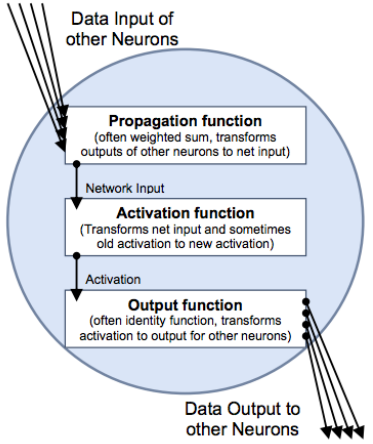
Neural nets derived as models of the human brain, in which each unit represents a neuron, and the connections represents synapses. The simplest model is called a Preceptron, and it consists on a single layer neural net with a series of inputs, one output, and a heavyside step activator function.



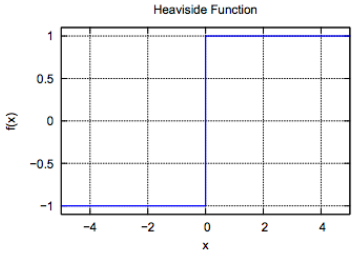
Definition A neural network is a graph (N, E, w) with a set of **neurons** (nodes/vertex) N, a set of **edges/connections** $E = (i, j)|i, j \in \mathbb{N}$, a set of **functions** $w : N \to \mathbb{R}$ that defines the weights of the connection from node i to node j, $w(i, j)$. Neural nets are extremely flexible and one has many choices available, mainly (1) the network structure/topology, for example, **feedforward networks** such that the neurons are grouped in the following layers: One input layer, n hidden processing layers (invisible from the outside, thats why the neurons are also referred to as hidden neurons) and one output layer

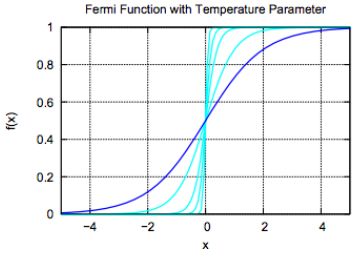


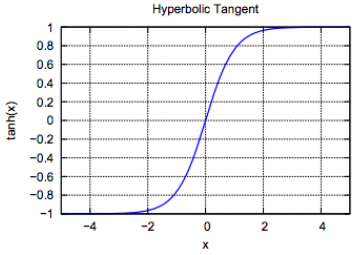
(2) the number of nodes and number of layers (complexity of the network),
(3) the propagation function, typically the weighted sum of neurons,



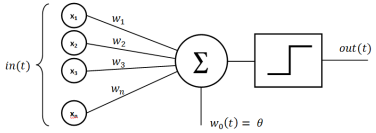
(4) and the activation/transfer function: step/heaviside, sigmoid/logistic, fermi with temperature parameter (smaller temperature close to heavyside) and the hyperbolic tangent.

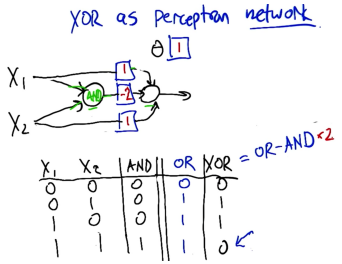






Preceptron The simplest neural net is the so-called preceptron: it consists of a set of inputs, a process neuron, and a single output - the activator is the step function. This simplest preceptron can output the AND, OR and NOT binary operations of two inputs, but not a XOR (not linearly separable).





Backpropagation A more general network is the