



Machine Learning

Background



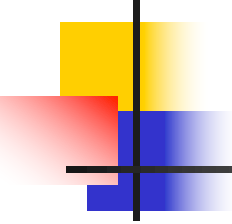
Machine Learning

- To perform machine learning there are generally three common, important elements
 - Representation
 - How the learning system represents/encodes the data and the learning results
 - Evaluation / Performance estimation
 - How the learning system determines the quality of what it has learned
 - Optimization
 - The mechanism that improves performance with more data



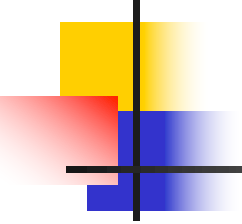
Representation and Hypothesis Space

- The representation used in the learning approach defines the hypothesis space
 - A hypothesis is a possible output of the learning system
 - The hypothesis space defines the space of all possible outputs of the system and can be defined in different terms
 - Parametrically through a set of parameters
 - E.g. the set of all 2nd order polygons for regression
 - ...
 - Non-parametrically through sets of data points



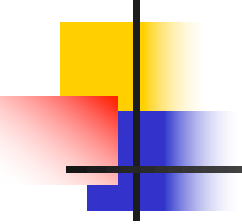
Hypothesis Space and Data Complexity

- The complete hypothesis space of a problem is usually enormous
 - In Boolean function approximation / classification it is exponential in the number of random variables
 - Every possible boolean function is one hypothesis
 - To learn the precise function requires generally to see all possible data
 - Learning the precise function is generally intractable



Hypothesis Space and Data Complexity

- To address the complexity of the hypothesis space, learning has to have a way to select among all the available solutions
 - Operate on a reduced hypothesis space
 - Removes possible solutions, leading to only approximate solutions
 - Use a preference criterion to select among possible hypotheses
 - Change the evaluation/performance/optimization function to distinguish/select between the different hypotheses.



Hypothesis Space and Data Complexity

- “Simplicity” of the hypothesis is one of the most commonly used criteria to select between hypotheses
 - E.g. minimum description length
 - Lowest degree polynomial function
 - Minimum number of hidden variables
 - Rationally usually based on either reduced complexity of subsequent use or Occam’s razor
- Other criteria include “closes” hypothesis, etc.



Evaluation

- The evaluation/performance measure determines what is to be learned from the data.
 - “Similarity
 - Error
 - Likelihood
 - Accuracy
 - Cost/Value
 - ...



Evaluation in Supervised Learning

- In supervised learning, the evaluation is largely based on the target output in the data
 - Evaluation expresses the quality of the hypothesis, e.g.
 - Error - Sum of squared differences between hypothesis and the target values over the data instances
 - Likelihood – Probability that the hypothesis will generate the target value for the data instances
 - Accuracy – (negative) Entropy of the hypothesis predictions in the set of data instances
 - Value – Utility of the hypothesis over the set of instances



Evaluation

- Based on the type of representation and evaluation used, machine learning make use of different fields
 - Probability Theory
 - E.g. maximum likelihood methods
 - Information Theory
 - E.g. minimum Entropy approaches
 - Utility Theory
 - E.g. maximum utility criteria



Probability and Statistics

- Probability and statistics are often used interchangeably but are different, related fields
 - Probability
 - Mainly a field of theoretical mathematics
 - Deals with predicting the likelihood of events given a set of assumptions (e.g. a distribution)
 - Statistics
 - Field of applied mathematics
 - Deals with the collection, analysis, and interpretation of data



Statistics

- Statistics deals with real world data
 - Data collection
 - How do we have to collect the data to get valid results
 - Analysis of data
 - What properties does the data have
 - What distribution does it come from
 - Interpretation of the data
 - Is it different from other data
 - What could cause issues in the data



Probability

- Probability is a formal framework to model likelihoods mainly used to make predictions
 - There are two main (interchangeable) views of probability
 - Subjective / uncertainty view: Probabilities summarize the effects of uncertainty on the state of knowledge

In Bayesian probability all types of uncertainty are combined in one number
 - Frequency view: Probabilities represent relative frequencies of events

$$P(e) = (\# \text{ of times of event } e) / (\# \text{ of events})$$



Probability

- Random variables define the entities of probability theory
 - Propositional random variables:
 - E.g.: IsRed, Earthquake
 - Multivalued random variables:
 - E.g.: Event, Color, Weather
 - Real-Valued random variables:
 - E.g.: Height, Weight



Axioms of Probability

- Probability follows a fixed set of rules
 - Propositional random variables:
 - $P(A) \in [0..1]$
 - $P(T) = 1, P(F) = 0$
 - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$
 - $P(A \wedge B) = P(A)P(B | A)$
 - $\sum_{x \in \{T, F\}} P(X = x) = 1$



Axioms of Probability

- The same axioms apply to multi-valued and continuous random variables
 - Multi-valued variables ($X \in S = \{x_1, \dots, x_N\}; Y, Z \subset S$):
 - $P(Y) \in [0..1]$
 - $P(Y = S) = 1, P(Y = \emptyset) = 0$
 - $P(Y \cup Z) = P(Y) + P(Z) - P(Y \cap Z)$
 - $P(Y \cap Z) = P(Y)P(Z|Y)$
 - $\sum_{x \in S} P(X = x) = 1$



Continuous Random Variables

- The probability of continuous random variables requires additional tools
 - The probability of a continuous random variable to take on a specific value is often 0 for all (or almost all) possible values
 - Probability has to be defined over ranges of values
 - Individual assignments to random variables have to be addressed using probability densities

$$p(X = x) \in [0.. \infty]$$



Continuous Random Variables

- Probability density is a measure of the increase in likelihood when adding the corresponding value to the range of values

$$P(a \leq X \leq b) = \int_a^b p(X = x) dx$$

- The probability density is effectively the derivative of the cumulative probability distribution

$$p(X = x) = \frac{dF(x)}{dx}; F(x) = P(-\infty \leq X \leq x)$$



Probability Syntax

- Unconditional or prior probabilities represent the state of knowledge before new observations or evidence
 - $P(H)$
- A probability distribution gives values for all possible assignments to a random variable
- A joint probability distribution gives values for all possible assignments to all random variables



Conditional Probability

- Conditional probabilities represent the probability after certain observations or facts have been considered
 - $P(H/E)$ is the posterior probability of H after evidence E is taken into account
 - Bayes rule allows to derive posterior probabilities from prior probabilities
 - $P(H \mid E) = P(E \mid H) P(H)/P(E)$



Conditional Probability

- Probability calculations can be conditioned by conditioning all terms
 - Often it is easier to find conditional probabilities
- Conditions can be removed by marginalization

- $$P(H) = \sum_E P(H | E)P(E)$$



Joint Distributions

- A joint distribution defines the probability values for all possible assignments to all random variables
 - Exponential in the number of random variables
 - Conditional probabilities can be computed from a joint probability distribution
 - $P(A | B) = P(A \cap B) / P(B)$



Inference

- Inference in probabilistic representation involves the computation of (conditional) probabilities from the available information
 - Most frequently the computation of a posterior probability $P(H/E)$ from a prior probability $P(H)$ and new evidence E



Statistics

- Statistics attempt to represent the important characteristics of a set of data items (or of a probability distribution) and the uncertainty contained in the set (or the distribution).
 - Statistics represent different attributes of the probability distribution represented by the data
 - Statistics are aimed at making it possible to analyze the data based on its important characteristics



Experiment and Sample Space

- A (random) experiment is a procedure that has a number of possible outcomes and it is not certain which one will occur
- The sample space is the set of all possible outcomes of an experiment (often denoted by S).
 - Examples:
 - Coin : $S=\{H, T\}$
 - Two coins: $S=\{HH, HT, TH, TT\}$
 - Lifetime of a system: $S=\{0..\infty\}$



Statistics

- A number of important statistics can be used to characterize a data set (or a population from which the data items are drawn)
 - Mean
 - Median
 - Mode
 - Variance
 - Standard deviation



Mean

- The arithmetic mean μ represents the average value of data set $\{X_i\}$

$$\mu = \frac{\sum_{i=1}^N X_i}{N}$$

- The arithmetic mean is the expected value of a random variable, i.e. the expected value of a data item drawn at random from a population

$$\mu = E[X]$$



Median and Mode

- The median m is the middle of a distribution

$$|\{X_i \mid X_i \leq m\}| = |\{X_i \mid X_i \geq m\}|$$

- The mode of a distribution is the most frequently (i.e. most likely) value



Variance and Standard Deviation

- The variance σ^2 represents the spread of a distribution

$$\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$$

- In a data set $\{X_i\}$ an unbiased estimate s^2 for the variance can be calculated as

$$s^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N-1}$$

- $N-1$ is often called the number of degrees of freedom of the data set
- The standard deviation σ is the square root of the variance
 - In the case of a sample set, s is often referred to as standard error



Moments

- Moments are important to characterize distributions
 - r^{th} moment: $E \left[(x - a)^r \right]$
 - Important moments:
 - Mean: $E \left[(x - 0)^1 \right]$
 - Variance: $E \left[(x - \mu)^2 \right]$
 - Skewness: $E \left[(x - \mu)^3 \right]$



Information Theory

- Information theory address the question how much information a particular event (or message) contains assuming no background knowledge
 - Information theory does not deal with the semantics (i.e. the meaning) of the event or the message.
 - It solely deals with the minimal amount of information required to represent its content in the absence of background information
 - How much information is required to perfectly remember the event/message ?
 - What is the minimal size to which the event/message can be compressed ?
- Information theory was established by Claude Shannon in the 1940s



Information Theory

- Information theory deals with a number of problems associated with information
 - How much information does an even provide us with ?
 - How far can we compress a message ?
 - What is the most compact encoding of a message ?
 - How much capacity does a transmission channel have if we use a particular encoding ?
- Information theory disregards any semantics of the message or knowledge outside the actual message
 - Information is directly tied to the probability of the event or message
 - Events that are easily predictable contain less information than ones that are more difficult to predict
 - Events that are known with certainty to occur do not provide any information since we already knew their content beforehand



Information

- The Information content, $I(p)$, of an event with likelihood p has to fulfill a number of properties
 - Information can not be negative
$$I(p) \geq 0$$
 - No occurrence of an event can take away from a the information in previous events since we are not modeling their semantics but rather are only looking at remembering / representing the events
 - An event that has probability 1, contains no information
$$I(1) = 0$$
 - We obtain no information from observing such an event and need no information to remember it.
 - The information provided by two independent events has to be the sum of the information from each one
$$I(p_1 * p_2) = I(p_1) + I(p_2)$$
 - The occurrence one event can not provide us with any information about the occurrence of an independent event.



Information

- Based on the required properties for a measure of information and assuming that it is continuous we can derive a measure for information:

$$I(p^x) = x * I(p) \quad \text{for all } 0 < p \leq 1, x > 0$$

$$0 \leq I(p)$$

$$\Rightarrow I(p) = -\log_b(p)$$

- The based b selects the unit in which we measure the information but is not important for any of the calculations
 - $b=2$: bits
 - $b=3$: trits
 - $b=e$: nats
 - $b=10$: Hartleys



Entropy

- Entropy represents the information content of a distribution
 - Entropy is the expected amount of information in a random sample from the distribution

$$\begin{aligned} H(X) &= -\sum_{x \in X} P(x) I(P(x)) \\ &= -\sum_{x \in X} P(x) \log_b(P(x)) \end{aligned}$$

- The best predictions are the ones that do not leave any entropy



Conditional Entropy

- The Conditional Entropy represents the incremental information beyond the condition

$$\begin{aligned} H(X | y) &= - \sum_{x \in X} P(x, y) I(P(x | y)) \\ &= - \sum_{x \in X} P(x, y) \log_b (P(x | y)) \end{aligned}$$

- This is akin to the posterior entropy
- Conditional entropy and (differential) entropy are related through

$$H(x, y) = H(x | y) + H(y)$$



Relative Entropy / Kullback-Leibler Divergence

- To address additional information needed when modeling a distribution P using a different distribution Q we use relative entropy

$$\begin{aligned} KL(P \parallel Q) &= -\sum_{x \in X} P(x) I(Q(x)) - \left(-\sum_{x \in X} P(x) I(P(x)) \right) \\ &= -\sum_{x \in X} P(x) \log_b \left(\frac{Q(x)}{P(x)} \right) \end{aligned}$$

- This is a measure of the dissimilarity of P and Q



Utility Theory

- Utility:
 - quantifies the degree of preference across different alternatives
 - models the impact of uncertainty on preferences
 - represents a mapping from an agent's state (or situation) to the degree of happiness (expected future payoff) for being in this state
- We will introduce this formally later



Evaluation/Performance

- Evaluation or Performance functions in machine learning algorithms can usually be defined in one of these frameworks
 - Probability theory is used for likelihood criteria
 - Information theory is usually used in accuracy/error rate criteria
 - Note that there is a strong link between entropy and probability and thus sometimes information theoretic criteria can be formulated as probabilistic criteria
 - All evaluation functions have to obey utility theory



Optimization

- The core of the learning algorithm is trying to improve (optimize) the value of the evaluation function
- Optimization can fall into two types depending on the available hypothesis space
 - Unconstrained optimization
 - All elements of the hypothesis space are allowed
 - Constrained optimization
 - Only some parts of the hypothesis space are allowed and the allowed part is defined by constraints



Optimization

- Different types of optimization techniques:
 - Global optimization
 - Analytic
 - Requires certain properties of the evaluation function, e.g.:
 - Continuous, differentiable almost everywhere
 - Absence of local minima (or only a finite number of them)
 - Uses the fact that the derivative of a function is 0 at local extrema
 - Incremental
 - Combinatorial search
 - Annealing techniques
 - ...
 - Global optimization is most of the time not tractable



Optimization

- Different types of optimization techniques:
 - Local optimization
 - Analytic
 - Requires certain properties of the evaluation function, e.g.:
 - Continuous, differentiable almost everywhere
 - Uses the fact that the derivative of a function is 0 at local extrema
 - Incremental
 - Gradient ascent/descent (requires continuity and differentiability)
 - ...
 - Addresses tractability issues but might not end up with the best solution



Optimization

- Different types of optimization techniques:
 - Local improvement
 - Incremental
 - Hill climbing
 - ...
 - Local improvement has the smallest number of requirements for the evaluation function but often progresses relatively slowly



Background

- Most machine learning techniques use elements from these theories and techniques to derive the algorithms
 - Throughout the course we will identify these for the learning approaches covered
 - Will go into more detail of the specific aspects of the formalism and the specific techniques then