CAPS Core Concepts

Leo Klenner, Henry Fung, Cory Combs

Last updated: 12/4/2019

Core Concepts

Concept	Description
Accuracy	*Metric* associated with *machine learning*. Accuracy measures percentages of total correct predictions (TP * TN) / (TP + FP + FN + TN). Common metric for balanced datasets but a poor metric for imbalanced datasets, when a large majority of outcomes are in one class rather than an other one
Action	Action refers to the output of an *agent*. Element of a *Markov decision process*
Agent	Algorithm built to make decisions or take *actions*. The agent's performance depends on specifications of the algorithm and the features of the *environment* in which the agent is situated. Element of a *Markov decision process*
AI	AI is the research field that aims to build systems that act rationally, where rationality is defined as *bounded optimality*
Algorithm	Finite sequence of instructions that can be read by a computer to solve a problem or perform a computation. Related to but different from a *model*
Adversarial example	*Risk* associated with *machine learning*. Data points included in the training dataset of a *machine learning algorithm* for malicious purposes to distort the algorithms performance. Especially relevant for *online learning* algorithms
Alignment	Concept often related to *reinforcement learning* focused on ensuring that the *actions* taken by an *agent* are aligned with human values. Ensuring alignment is challenging because of failure modes like *reward gaming*. Related to *human-in-the-loop learning*
AUC-ROC curve	Short for area under the curve receiver operating characteristic curve. Statistic used for *model* comparison to find an optimal point between the *sensitivity* and *specificity* of a *model*

Concept	Description
Audit	To uncover and mitigate *risk* in *machine learning models* and *algorithms*, audits need to be performed. Differentiate between audits in cooperative settings, where access is granted to the code, and non-cooperative settings, where no access is granted to the code. In non-cooperative settings, audit tools are often limited to *black box testing* and audit results need to be interpreted carefully to avoid introducing confirmation *bias*
Bagging	See *bootstrap aggregation*
Bias	Both a common trade-off and an ethical *risk* associated with *machine learning*. Bias refers to the difference between an estimator's expected value and the true value of the *parameter* being estimated. Bias can arise from non-representative data, poorly constructed algorithms and human interpretation of results, leading to "garbage in, garbage out", which we focus on in the ethics context. See also *variance*, and *bias-variance trade-off*
Bias-variance trade-off	The bias-variance trade-off is a property of predictive *machine-learning models* whereby models with lower *bias* have higher *variance* and vice versa. Example: a complex nonlinear model that fits the training data very well has low *bias* but high *variance* as the model will change substantially if different training data is used. Resolved through *bootstrap aggregation* and *gradient boosting*
Binary classification	Type of *classification* to classify two possible outputs, frequently represented numerically as 0 and 1. Often implemented through *logistic regression*. Different from multiclass classification
Black box testing	Method in software testing that threats a *model* or *algorithm* as a black-box and aims to test the algorithm through an analysis and comparison of input-output pairs. Inputs are provided, outputs are obtained. Primarily used to identify what a *model* or *algorithm* does, not how. Does not require access to code and often used in *audits*, especially those in non-cooperative settings
Boosting	See *gradient boosting*
Bootstrap aggregation	Bootstrap aggregation, also known as *bagging*, is a technique that can improve the predictive *accuracy* of a number of *machine learning* methods. It's designed to reduce the *variance* of large *decision-tree models* that have low *bias*. The main idea is that we bootstrap multiple training datasets, through sampling with replacement from a small dataset, and then aggregate the predictions for each of these datasets to generate a low *variance* prediction. See also *bias-variance trade-off*

Concept	Description
Bounded optimality	Type of rational action; actions are optimal with the bounds of available *computing power*, information, and time horizons imposed on decisions
Chi-square	*Algorithm* part of *decision tree models*. Chi-square measures the statistical significance of a given data split, allowing the *model* to choose the maximally significant split
Classification	Branch of *supervised learning*. In classification the *model* predicts a categorical target variable. Different from *regression*
Complex	The property that components of a system interact in ways that cannot be described through higher instructions, such as *rules*. Associated with *non-deterministic* and different from *complicated*. Note, that complexity in computer science often has a different meaning: run-time complexity related to *execution*
Complicated	The property of having multiple features or *rules* that, in their sum, can be difficult to understand. Different from *complex*
Confusion matrix	Table that anchors the essential *metrics* for *classification models*
Constraint	Two meanings (1) constraints whether an *algorithm* can solve a problem, ie. the number of *computing power* available; (2) constraints that determine how an *algorithm* can solve a
Computing power	Hardware aspects that determine an algorithm's performance, ie. its speed. Also referred to as computational resources or compute
Cross-validation	Technique to split training data for *supervised learning*. Cross-validation splits the data into a specified number of different training and tests sets; trains and tests the *model* using each split; and compares the results of each case. Smoothes the potential for *bias* in any given random selection
Data oversight	Lack thereof is a *risk* associated with *machine learning*. Includes *adversarial examples*, *bias*, *overfitting*, *spuriousness*
Decision tree	*Model* for *supervised learning* that can handle both *classification* and *regression*. Contains nodes, branches, and leaves. The basic intuition of a decision tree is similar to the game "twenty questions", in which a series of yes/no questions leads to a prediction of the hidden answer. The decision tree formalizes the process of questioning into a reproducible analysis of specific *features*, while allowing for a far broader range of questions.

Concept	Description
Design specification	Stage of a system's *specification*. The blueprint through wish a system is specified. Preceded by *design specification* and precedes *revealed specification*
Deterministic	The property of an *algorithm* that for the same input, it will always return the output. Associated with *rules-based systems*, different from *non-deterministic*
Domain knowledge	Expert knowledge about a specific *environment*, like a region or a subject matter like strategy. Needed for *rule-based systems* and *robust* algorithmic decision-making
Ensemble models	Aggregation of *decision-trees*, ie. into *random forests*. Idea is to average results across multiple decision trees, thus smoothing potential *overfitting* by each *model*
Environment	Space where the decisions or actions of the agent are carried out and whose features, like partial observability, impact the agent's performance. Element of a *Markov decision process*
Epsilon-greedy	A type of *greedy algorithm*. Provides a simple solution to the *explore-exploit dilemma*. An epsilon greedy algorithm explores a random *action* with probability epsilon and exploits the known best action with probability 1-epsilon.
Execution	Process by which a program executes instructions, commonly referred to as "running the program"
Explainability	Lack thereof is a *risk* associated with *machine learning*. Given the n*on-deterministic* nature of *machine learning*, it isn't always clear why an *algorithm* arrived at a certain result. Even if there is clarity in a strictly mathematical sense, this perspective might still not be understandable for the average end-user. Hence, a subset of research actively explores explainable *AI*
Explore-exploit dilemma	Trade-off associated with *optimization*. Should the *agent* exploit the current best optimum to maximize short-term returns and risk being stuck in a *local optimum*? Or should the agent explore the *search space* and forego the returns from the current best optimum to find the *global optimum* and thus maximize long-term returns?
Feature	The variables in a dataset (the columns in a table). Related to *observations*
Feature selection	Process of selecting the right *features* to train a *supervised learning model*. Often draws heavily on qualitative reasoning and trial-and-error

Concept	Description
Gini index	*Algorithm* part of *decision tree models*. The Gini index measures the probability that a random *classification* will be incorrect.
Global optima	Optimal points in a *search space* that are optimal for the entire space. Associated with *optimization* and different from *local optima*
Gradient	Gradient is a term for derivative, or the rate of change of a function
Gradient boosting	Gradient boosting, also known as *boosting*, is a technique to improve the predictive *accuracy* of *machine learning models*. The main idea is that boosted models can "learn from their past errors". In boosting, a series of models are built where each model takes into account the error from the previous model. Thus, error is reduced sequentially. Boosted models are hard to interpret and lack *transparency*. See also *bias-variance trade-off*
Greedy	Class of algorithms. Greedy algorithms *exploit* the *local optimum* of a decision space and often cannot discover the *global optimum*. Example of a greedy algorithm is *epsilon-greedy*
Hyperparameter	Variable that is external to a *model* and whose value must be tuned in order to obtain a model with optimal performance. Related to but different from *parameters*
Human-in-the-loop learning	Technique often related to *reinforcement learning* through which a human retains control over the *actions* an *agent* selects as it learns. Actions that are deemed undesirable are blocked. Often draws on *domain knowledge* and aims to ensure *alignment*
Ideal specification	Stage of a system's *specification*. The wishes through which a system is specified. Precedes *design specification*
Information gain	*Algorithm* part of *decision tree models*. Information gain determines which *features* yield the most information, using a concept called entropy
Local optima	Optimal points in a *search space* that are optimal only for a sub-region of the space. Associated with *optimization* and different from *global optima*
Logistic regression	*Model* used as the leading means for *binary classification*. Logistic regression employs a logistic function to transform an array of possible values into a value close to 1 or 0. The function produces a probabilistic output, and we can decide a probability cut-off, known as the decision boundary, to classify values as 1 or 0.

Concept	Description
Machine learning	One of two main types of *AI* together with *rules-based systems*. Machine learning *algorithms* train on large datasets, yield *non-deterministic* decision-making, and are *complex*. Subsets of machine learning include *supervised learning*, *unsupervised learning*, *meta learning*, and *transfer learning*
Markov decision process	Often abbreviated as MDP. Framework for formalizing problems to enable the application of *reinforcement learning algorithms* to them. In a MDP, an *agent* interacts with an *environment* and for each *action*, receives a *reward* and transitions into a new *environment state*. Assumes that the *Markov property* holds
Markov property	The Markov property states that the future is independent of the past given the present. This means that, given the information available to us in the present, knowing the past would not enable us to make better predictions about the future. Assumed in *Markov decision processes*
Meta learning	Type of *machine learning*. Meta learning refers to teaching *algorithms* learning how to learn and is a current frontier in *AI* research. Related to but conceptually distinct from *transfer learning*
Metrics	Used to measure essential properties of *machine learning models*. Encompasses *accuracy*, *precision*, *sensitivity*, *specificity*
Model	A model refers to an *algorithm* that was trained on data
Model development	Workflow that encompasses data preparation (split the data into training and test datasets), model construction (build the model), model training (fit the model to the training data), model testing (predict values using the test data) and model deployment (evaluate the model's performance)
Multi-armed bandit	Computational model of decision-making under uncertainty related to *reinforcement learning*. In a multi-armed bandit, an *agent* has to learn the rewards for a number of actions through trial-and-error sampling of these actions. Presents an *online learning* problem and incorporates the *explore-exploit dilemma*. Can be formalized as a *Markov decision process* with only one *state*. Multi-armed bandits are often used to model recommender problems. Most basic solution to multi-armed bandits is the *epsilon-greedy algorithm*
Non-deterministic	The property of an algorithm that for the same input, it will not return the output. Associated with *machine learning*, different from *deterministic*
Observation	The number of unique records (rows in a table). Related to *features*

Concept	Description
Offline learning	Algorithms that train on a dataset prior to their deployment and do not learn in real-time during their deployment. Advantages are high predictability of performance, disadvantages are lower adaptability. Different from *online learning*
Online learning	Algorithms that train in real-time during their deployment. Advantages are high adaptability, disadvantages are low predictability of performance. Different from *offline learning*
Optimization	*Machine learning* problems can be expressed as mathematical optimization problems in which we aim to minimize the rate or errors or maximize the rate of success for a task. Optimization is best conceptualized as a *search* across a continuous *search space* that contains the actions the agent can perform. Optimization can return *local optima* (undesired) and *global optima* (desired). To get from the former to the latter, we often add *constraints* to the optimization
Overconfidence	*Risk* associated with *machine learning*. Humans might be prone to be overconfident in the results of *algorithms*, for a variety of reasons. Can lead to an amplification of risks from *spuriousness* and others
Overfitting	*Risk* associated with *machine learning*. If presented with a training set that is too large, an *algorithm* might learn oversimplified behavior from the data that doesn't generalize to other datasets and thus leads to suboptimal performance
Parameter	Configuration variable that is internal to a *model* and whose value must be estimated using the training dataset. Related to but different from *hyperparameters*
Policy	Term associated with *reinforcement learning* and not to be confused with policy in the sense of governance. In *reinforcement learning*, a policy is a mapping from *states* to *actions*. The *agent* has to learn an optimal *policy* to achieve a task.
Precision	*Metric* associated with *machine learning*. Precision measures the percentage of predicted positives that were correct (TP) / (TP + FP)
Pruning	Technique used in building *decision-tree models* to remove sections of the tree that provide little predictive power and that could, if left unchecked, lead to *overfitting*
Random forest	*Ensemble method* in *supervised learning* that operates by constructing a multitude of *decision-trees* at training time. Random forests can be created through *bootstrap aggregation* and have higher predictive accuracy than decision-trees or linear regression. Random forests are hard to interpret and require a lot of *computing power*

Concept	Description
Regression	Branch of *supervised learning*. In regression, the *model* predicts a continuous (numerical) target variable. Different from *classification*
Reinforcement learning	Type of *machine learning*. Reinforcement learning is about learning through trial-and-error from interaction with an *environment* what the best sequence of actions is to achieve a specified goal, ie. for autonomous control
Revealed specification	Stage of a system's *specification*. The specification that best describes the system's actual behavior. Becomes a *risk* if unaligned with the system's *designed specification*. Preceded by *ideal specification* and *design specification*
Reward	Numerical signal sent to the *agent* through the *environment* to provide feedback on *actions* taken in a *state*. In *reinforcement learning*, qualitative goals are decomposed into a numerical reward signal which guides the *agent* and enables it to learn an optimal *policy*. Rewards are issued to the agent through human-designed reward functions and poorly designed reward functions can lead to failure modes, such as *reward gaming*. Element of a *Markov decision process*
Reward gaming	Failure mode in *reinforcement learning*. An *agent* learns to maximize *rewards* in a manner that does not represent the intention of its designers. Reward gaming constitutes a disconnect between *desired specification* and *revealed specification*
Risk	Manifold types of risks are associated with *AI*. Many of these risks arise from how machines interact with humans. In session 1, we have focused on safety risks, but there are others, such as governance risks
Robustness	Goal when designing an *agent* is to enable robust decision-making that can cope with errors during execution
Rules	If-then-else statements, like "if condition A is True, then do action B, else do action C" used to build *rule-based systems* based on *domain knowledge*
Rule-based systems	One of two main types of *AI* together with *machine learning*. Algorithm constructed from *rules*, yields *deterministic* decision-making. Can be *complicated* but not *complex*
Search space	Spatial representation of possible solution points to a problem, from which a *machine learning algorithm* should learn to select the *global optimum*. Associated with *optimization*
Sensitivity	*Metric* associated with *machine learning*. Sensitivity measures percentage of actual positives that were correctly predicted: (TP) / (TP + FN)

Concept	Description
Sigmoid function	Basic logistic function for *binary classification*
Specification	Mathematical properties that define the implementation of a system. Stages include *ideal specification*, *design specification*, *revealed specification*
Specificity	*Metric* associated with *machine learning*. Specificity measures percentage of actual negatives that were correctly predicted: (TN)/(TN + FP)
Spuriousness	*Risk* associated with *machine learning*. Results are spurious if they are identified based on an artificial correlation between unrelated covariates. Colloquially, "the data tricks the *algorithm* into seeing something that isn't there". Especially relevant in high-dimensional datasets
State	State, here treated as an *environment state*, describes the subsets through which an *environment* can be exhaustively represented. This representation is arrived at through human choice, ie. we can choose to represent the world (environment) as land and sea (states) or different countries (states) or as either B.C. or A.C. (states). Element of *Markov decision processes*
Supervised learning	Type of *machine learning*. Supervised learning is learning from labeled data, where the labels are provided to the algorithm by a human and the algorithm learns to apply these labels to new data, ie. for prediction
Transfer learning	Type of *machine learning*. Transfer learning refers to teaching *algorithms* how to generalize knowledge learned in one domain to another domain. This can include generalizations from games to logistics, or from virtual environments to physical environments. Frontier in current *AI* research. Related to but conceptually distinct from *meta learning*
Transparency	Lack thereof is a *risk* associated with *machine learning*. Given the non-deterministic nature of *machine learning algorithms* there decision-making is often opaque and presents a "black box" that humans can't access
Unsupervised learning	Type of *machine learning*. Unsupervised learning is learning from data that does not have labels, for the purpose of identifying features that can be used to group or cluster the data, ie. for pattern detection
Variance	Variance refers to how far a set of random numbers are spread out from their average value. See also *bias* and *bias-variance trade-off*