Glossary

ACRE-learnable: The original framework proposed by Lowd & Meek (2005*a*) for quantifying the query complexity of a family of classifiers; see also, near-optimal evasion problem. See 49.

action: In the context of a learning algorithm, a response or decision made by the learner based on its predicted state of the system. See 20, 21, 27.

adversarial learning: Any learning problem where the learning agent faces an adversarial opponent that wants the learner to fail according to a well-defined adversarial objective. Specifically, in this book, we consider adversarial learning in security-sensitive domains. See 18, 56, 104.

anomaly detection: The task of identifying anomalies within a set of data. See 4, 8, 18, 26, 28, 37, 41, 52, 69, 70, 134–138, 144, 200, 243.

approximate optimality: A notion of optimality in which a valid assignment achieves a value that is close to the optimal achievable value for a particular optimization problem. The notion of *closeness* can be defined in several ways.

▶ **additive gap** ($G^{(+)}$): The additive difference between the estimated optimum $\hat{C}$ and the global optimum C^* as measured by the difference between these two quantities: $\hat{C} - C^*$. When the global optimum is not known, this gap refers to the difference between the estimated optimum and a *lower bound* on the global optimum. See 207.

▶ **additive optimality**: A form of approximate optimality where the estimated optimum $\hat{C}$ is compared to the global optimum C^* using the difference $\hat{C} - C^*$; η-additive optimality is achieved when this difference is less than or equal to η. See 206–208, 212, 217, 225.

▶ **multiplicative gap** ($G^{(*)}$): The multiplicative difference between the estimated optimum $\hat{C}$ and the global optimum C^* as measured by the ratio between these two quantities: $\frac{\hat{C}}{C^*}$. When the global optimum is not known, this gap refers to the ratio between the estimated optimum and a *lower bound* on the global optimum. See 207, 213, 220.

▶ **multiplicative optimality**: A form of approximate optimality where the estimated optimum $\hat{C}$ is compared to the global optimum C^* using the ratio $\frac{\hat{C}}{C^*}$; ϵ-multiplicative

https://doi.org/10.1017/9781107338548.014 Published online by Cambridge University Press

optimality is achieved when this ratio is less than or equal to $1 + \epsilon$. See 206–208, 211, 212, 217, 227–230.

attacker: In the learning games introduced in Chapter 3, the attacker is the malicious player who is trying to defeat the learner. See 26, 36.

batch learning: A learning process in which all training data is examined in batch by the learning algorithm to select its hypothesis, f. See 25, 50.

beta distribution: A continuous probability distribution with support on (0, 1) parameterized by $\alpha \in \Re_+$ and $\beta \in \Re_+$ that has a probability density function given by $P(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}$. See 109, 110.

beta function $(\mathrm{B}(\alpha, \beta))$: A two-parameter function defined by the definite integral $\mathrm{B}(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\ dt$ for parameters $\alpha > 0$ and $\beta > 0$. See 109.

blind spot: a class of miscreant activity that fails to be correctly detected by a detector; i.e., false positives. See 18, 199.

boiling frog poisoning attack: An episodic poisoning method spanning several training iterations, which is named after the folk tale that one can boil a frog by slowly increasing the water temperature over time. In a boiling frog attack, the adversary increases the total amount of poisoned data used during each subsequent training step according to some poisoning schedule so that the detector is gradually acclimated to this malicious data and fails to adequately identify the eventually large quantity of poisoning that has been introduced. See 144, 151, 159–163.

breakdown point $(\epsilon^\star)$: Informally, it is the largest fraction of malicious data that an estimator can tolerate before the adversary can use the malicious data to arbitrarily change the estimator. The breakdown point of a procedure is one measure of its robustness. See 56, 57, 145, 147, 249.

chaff: Extraneous noise added into a data source to mislead a detector. In the case of PCA-based network anomaly detection in Chapter 6, chaff is spurious traffic sent across the network by compromised nodes to interfere with PCA's subspace estimation procedure. See 134, 136, 140.

classification: A learning problem in which the learner is tasked with predicting a response in its response space $\mathcal{Y}$ given an input x from its input space $\mathcal{X}$. In a classification problem, the learned hypothesis is referred to as a classifier. The common case when the response case is boolean or $\{0, 1\}$ is referred to as binary classification. See 23, 26.

▶ **binary classification**: A classification learning problem where the response space $\mathcal{Y}$ is a set of only two elements; e.g., $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{"+", "-"\}$. See 26, 28.

classifier (f): A function $f : \mathcal{X} \to \mathcal{Y}$ that predicts a response variable based on a data point $\mathbf{x} \in \mathcal{X}$. In classification, the classifier is selected from the space $\mathcal{F}$ based on a

https://doi.org/10.1017/9781107338548.014 Published online by Cambridge University Press

labeled dataset $\mathbb{D}^{(\text{train})}$; e.g., in the empirical risk minimization framework. See 26, *see also* hypothesis.

convex combination: A linear combination $\sum_i \alpha_i \cdot \mathbf{x}^{(i)}$ of the vectors $\left\{\mathbf{x}^{(i)}\right\}$ where the coefficients satisfy $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. See 76, 108, 257.

convex function: A real-valued function $g : \mathbb{X} \rightarrow \Re$, whose domain $\mathbb{X}$ is a convex set in a vector space, is a convex function if for any $x^{(1)}, x^{(2)} \in \mathbb{X}$, the function satisfies the inequality

$$g\left(\alpha x^{(1)} + (1-\alpha)x^{(2)}\right) \leq \alpha g\left(x^{(1)}\right) + (1-\alpha)g\left(x^{(2)}\right) \quad ,$$

for any $\alpha \in [0, 1]$. See 177, 179–182, 184, 186–189, 192, 194, 210–212, 216, 221, 231, 259.

convex hull: The smallest convex set containing the set $\mathbb{X}$, or equivalently, the intersection of all convex sets containing $\mathbb{X}$, or the set of all convex combinations of the points in $\mathbb{X}$. For a finite set $\mathbb{X} = \left\{x^{(i)}\right\}$, its convex hull is thus given by $\mathbb{C}_{\mathbb{X}} = \left\{\sum_i \alpha_i x^{(i)} \;\middle|\; \sum_i \alpha_i = 1 \;\wedge\; \forall\, i\; \alpha_i \geq 0\right\}$. See 213, 214, 227, 228.

convex optimization: The process of minimizing a convex function over a convex set. See 94, 147, 177, 184, 195, 211, 232.

convex set: A set $\mathbb{A}$ is convex if for any pair of objects $a, b \in \mathbb{A}$, all convex combinations of a and b are also in $\mathbb{A}$; i.e., $\alpha a + (1-\alpha) b \in \mathbb{A}$ for all $\alpha \in [0, 1]$. See 200, 201, 210–212, 215, 218, 221–225, 231, 232, 257, 259.

convex-inducing classifier: A binary classifier f for which either $\mathcal{X}_f^+$ or $\mathcal{X}_f^-$ is a convex set. See 11, 18, 200–202, 206, 209–211, 218, 225, 231, 232, 244, 245, *see also* classifier.

cost function: A function that describes the cost incurred in a game by a player (the adversary or learner) for its actions. In this book, the cost for the learner is a loss function based solely on the learner's predictions whereas the cost for the adversary may also be data dependent. See 32.

***K*-covering of $\mathbb{X}$**: A collection of K balls of size ϵ (i.e., sets $\mathbb{B}_i$) arranged such that the object represented by set $\mathbb{X}$ is completely contained within their union; i.e., $\mathbb{X} \subseteq \bigcup_i \mathbb{B}_i$.

▶ **covering number**: The minimum number of balls needed to cover an object and hence, a measure of the object's complexity. See 205, 230.

data: A set of observations about the state of a system. See 20, *see also* dataset.

data collection: The process of collecting a set of observations about the system that comprise a dataset. See 23, 45.

data point ($\mathbf{x}$): An element of a dataset that is a member of $\mathcal{X}$. See 23, 26, 27.

data sanitization: The process of removing anomalous data from a dataset prior to training on it. See 18.

https://doi.org/10.1017/9781107338548.014 Published online by Cambridge University Press

dataset ($\mathbb{D}$): An indexed set of data points denoted by $\mathbb{D}$. See 23, 25, 26.

deep neural network: Multilayer neural network models that use cascades of hidden layers to implicitly undertake complex tasks such as feature extraction and transformation as part of the learning process. See 44, 48.

defender: In the learning games introduced in Chapter 3, the defender is a learning agent that plays against an attacker. If the learning agent is able to achieve its security goals in the game, it has achieved secure learning. See 26, 36.

degree of security: The level of security expected against an adversary with a certain set of objectives, capabilities, and incentives based on a threat model. See 13.

denial-of-service (DoS) attack: An attack that disrupts normal activity within a system. See 9, 18, 44, 52, 134, 138, 140–144, 146, 151–153, 163, 243.

dictionary attack: A *Causative Availability* attack against SpamBayes, in which attack messages contain an entire dictionary of tokens to be corrupted. See 112, 113.

differential privacy: A formal semantic information-theoretic measure of the level of training dataset privacy preserved by a learner publicly releasing predictions. See 18, 30, 62, 64–66, 171, 173–176, 179, 182, 183, 185, 186, 197, 198.

dispersion: The notion of the spread or variance of a random variable (also known as the scale or deviation). Common estimators of dispersion include the standard deviation and the median absolute deviation. See 136, 145, 147, 148.

distributional robustness: A notion of robustness against deviations from the distribution assumed by a statistical model; e.g., outliers. See 136.

DNN: See 44, 48, *Glossary:* deep neural network.

empirical risk ($\tilde{R}_N(f)$): The average loss of a decision procedure f with respect to data from a dataset $\mathbb{D}$. See 26, 57, 176, 188, 189.

empirical risk minimization: The learning principle of selecting a hypothesis that minimizes the empirical risk over the training data. See 22, 25, 26, 57, 109, 112, 176, 197, *see also* risk.

Erlang q-distribution: A continuous probability distribution with support on $[0, \infty)$ parameterized by a shape $q \in \mathfrak{N}$ and a rate $\lambda \in \mathfrak{R}_+$ that has a probability density function given by $P(x) = \frac{x^{q-1}\exp(-x/\lambda)}{\lambda^q(q-1)!}$. See 182, 183, 188.

expert: An agent that can make predictions or give advice that is used to create a composite predictor based on the advice received from a set of experts. See 56, 58–60, 133, 249, 250, 252.

explanatory variable: An observed quantity that is used to predict an unobservable response variable. See 23, *see also* data point.

false negative: An erroneous prediction that a positive instance is negative. See 27, 31, 32, 34, 36, 51, 114, 115, 119, 134, 156, 157.

false positive: An erroneous prediction that a negative instance is positive. See 27, 32, 34, 35, 71, 114–117, 156, 157.

false-negative rate: The frequency at which a predictor makes false negatives. In machine learning and statistics, this is a common performance measure for assessing a predictor along with the false-positive rate. See 27, 52, 96, 105, 120, 134, 137, 151, 157, 162, *see also* false negative.

false-positive rate: The frequency at which a predictor makes false positives. In machine learning and statistics, this is a common performance measure for assessing a predictor along with the false-negative rate. See 27, 28, 102, 105, 109, 120, 131, 134, 151, 157, 162, 243, *see also* false positive.

feature: An element of a data point; typically a particular measurement of the overall object that the data point represents. See 22, 30, 31, 38, 46, 47, 52, 63, 113, 212, 213, 219, 234, 246.

feature deletion attack: An attack proposed by Globerson & Roweis (2006) in which the adversary first causes a learning agent to associate intrusion instances with irrelevant features and subsequently removes these spurious features from its intrusion instances to evade detection. See 41, 46.

feature selection: The second phase of data collection in which the data are mapped to an alternative space $\hat{\mathcal{X}}$ to select the most relevant representation of the data for the learning task. In this book, we do not distinguish between the feature selection and measurement phases; instead they are considered to be a single step, and $\mathcal{X}$ is used in place of $\hat{\mathcal{X}}$. See 24, 30, 31, 41, 46, 47, 246, 247.

feature selection map (ϕ): The (data-dependent) function used by feature selection to map from the original input space $\mathcal{X}$ to a second feature space $\hat{\mathcal{X}}$ of the features most relevant for the subsequent learning task. See 21, 24, 46, 72, 236, 247.

Gaussian distribution ($\mathrm{N}(\boldsymbol{\mu}, \sigma)$): A (multivariate) continuous probability distribution with support on $\Re^D$ parameterized by a center $\boldsymbol{\mu} \in \Re^D$ and a scale $\sigma \in \Re_+$ that has a probability density function given by $P(\mathbf{x}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|\mathbf{x}-\boldsymbol{\mu}\|_2^2}{2\sigma^2}\right)$. See 25, 136, 159, 160, 235.

good word attack: A spam attack studied by Wittel & Wu (2004) and Lowd & Meek (2005*b*), in which the spammer adds words associated with non-spam messages to its spam to evade a spam filter. More generally, any attack where an adversary adds features to make intrusion instances appear to be normal instances. See 41, 42.

gross-error model ($\mathcal{P}_\epsilon(F_\mathcal{Z})$): A family of distributions about the known distribution $F_\mathcal{Z}$ parameterized by the fraction of contamination ϵ that combine $F_\mathcal{Z}$ with a fraction ϵ of contamination from distributions $H_\mathcal{Z} \in \mathcal{P}_\mathcal{Z}$. See 57.

https://doi.org/10.1017/9781107338548.014 Published online by Cambridge University Press

gross-error sensitivity: The supremum, or smallest upper bound, on the magnitude of the influence function for an estimator; this serves as a quantitative measure of a procedure or estimator. See 58.

Hilbert space ($\mathcal{H}$): An inner-product space that is complete with respect to the metric induced by its inner product; i.e., the metric $\|\mathbf{x}\|_{\mathcal{H}} \triangleq \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ for all $\mathbf{x} \in \mathcal{H}$. See 102, 178, 188.

▶ **reproducing kernel Hilbert space**: A Hilbert space $\mathcal{H}$ of real-valued functions on the space $\mathcal{X}$, which includes, for each point $\mathbf{x} \in \mathcal{X}$, a point-evaluation function $k(\cdot, \mathbf{x})$ having the reproducing kernel property $\langle f, k(\cdot, \mathbf{x}) \rangle_{\mathcal{H}} = f(\mathbf{x})$ for all $f \in \mathcal{H}$. See 177–179, 185, 186, 188, 189, 197, 198, 233.

hypothesis (f): A function f mapping from the data space $\mathcal{X}$ to the response space $\mathcal{Y}$. The task for a learner is to select a hypothesis from its hypothesis space to best predict the response variables based on the input variables. See 21, 22, 24–27, *see also* classifier.

hypothesis space ($\mathcal{F}$): The set of all possible hypotheses, f, that are supported by the learning model. While this space is often infinite, it is indexed by a parameter $\boldsymbol{\theta}$ that maps to each hypothesis in the space. See 24–26.

IDS: See 31, 40, 41, 45, 47, 53, 201, *Glossary:* intrusion detection system.

index set: A set $\mathbb{I}$ that is used as an index to the members of another set $\mathbb{X}$ such that there is a mapping from each element of $\mathbb{I}$ to a unique element of $\mathbb{X}$. See 256.

indicator function: The function $\mathrm{I}[\cdot]$ that is 1 when its argument is true and is 0 otherwise. See 256.

inductive bias: A set of (implicit) assumptions used in inductive learning to bias generalizations from a set of observations. See 22, 24.

inductive learning: A task where the learner generalizes a pattern from training examples; e.g., finding a linear combination of features that empirically discriminates between positive and negative data points. See 22.

influence function ($\mathrm{IF}(z; H, F_{\mathcal{Z}})$): A functional used extensively in robust statistics that quantifies the impact of an infinitesimal point contamination at z on an asymptotic estimator H on distribution $F_{\mathcal{Z}}$; see Section 3.5.4.3. See 56, 57, 249.

input space ($\mathcal{X}$): The space of all data points. See 23, *see also* data point.

intrusion detection system: A detector that is designed to identify suspicious activity that is indicative of illegitimate intrusions. Typically these systems are either host-based or network-based detectors. See 31.

intrusion instance: A data point that corresponds to an illegitimate activity. The goal of malfeasance detection is to properly identify normal and intrusion instances and prevent the intrusion instances from achieving their intended objective. See 26.

https://doi.org/10.1017/9781107338548.014 Published online by Cambridge University Press

intrusion prevention system: A system tasked with detecting intrusions and taking automatic actions to prevent detected intrusions from succeeding. See 31, *see also* intrusion detection system.

iterated game: In game theory, a game in which players choose moves in a series of repetitions of the game. See 58, 71.

kernel function (k): A bivariate real-valued function on the space $\mathcal{X} \times \mathcal{X}$, which implicitly represents an inner product in some Hilbert space so long as it is a positive semi-definite function. See 69, 102, 177–180, 182, 186, 187, 189, 191, 192, 197, 198.

> ▶ **RBF kernel**: A commonly used kernel function for numeric-valued data. The radial basis function (RBF) kernel is defined here as $k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \exp\left(-\frac{\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|_2^2}{2\sigma^2}\right)$ where $\sigma > 0$ is the kernel's bandwidth parameter. See 174, 182, 197, *see also* Gaussian distribution.

label: A special aspect of the world that is to be predicted in a classification problem or past examples of this quantity associated with a set of data points that are jointly used to train the predictor. See 23, 25–28, 33, 55, 58, 59, 63, 106, 107, 109, 110, 114, 115, 120, 130, 201, 235.

labeled dataset: A dataset in which each data point has an associated label. See 23.

Laplace distribution ($Laplace(\boldsymbol{\mu}, \lambda)$): A (multivariate) continuous probability distribution with support on $\Re^D$ parameterized by a center $\boldsymbol{\mu} \in \Re^D$ and a scale $\lambda \in \Re_+$ that has a probability density function given by $P(\mathbf{x}) = \frac{1}{2\lambda} \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}\|_1}{\lambda}\right)$. See 173, 182.

Laplace noise: A random variable drawn from a **0**-centered Laplace distribution and added to a nonrandom quantity. Laplace noise is used extensively in Chapter 7 through the Laplace mechanism to provide privacy properties to nonprivate learners. See 173, 179, 186, 188, 198, 244, *see also* Laplace distribution.

learner: An agent or algorithm that performs actions or makes predictions based on past experiences or examples of how to properly perform its task. When presented with new examples, the learner should adapt according to a measure of its performance. See 24.

learning algorithm: Any algorithm that adapts to a task based on past experiences of the task and a performance measure to assess its mistakes. See 25.

loss function (ℓ): A function, commonly used in statistical learning, that assesses the penalty incurred by a learner for making a particular prediction/action compared to the best or correct one according to the *true* state of the world; e.g., the squared loss for real-valued prediction is given by $\ell(y, \hat{y}) \triangleq (\hat{y} - y)^2$. See 21, 25–27, 40, 50, 58, 59, 111, 112, 174, 177, 179, 180, 182–184, 186, 187, 192.

https://doi.org/10.1017/9781107338548.014 Published online by Cambridge University Press

machine learning: A scientific discipline that investigates algorithms that adapt their behavior based on past experiences and observations. As stated by Mitchell (1997), "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E." See 20.

malfeasance detection: The task of detecting some particular form of illegitimate activity; e.g., virus, spam, intrusion, or fraud detection. See 4.

measurement: An object mapped from the space of real-world object to the data representation used by a learning algorithm. See 22.

measurement map: A description of the process that creates a measurement based on the observations and properties of a real-world object. See 23.

median absolute deviation: A robust estimator for dispersion defined by Equation (6.5), which attains the highest possible breakdown point of 50% and is the most robust M-estimator for dispersion. See 136.

membership query: A query sent to an oracle to determine set membership for some set defined by the oracle's responses. See 203, 231, 245.

mimicry attack: An attack where the attacker tries to disguise malicious activity to appear to be normal. See 41, 47, 54, 201.

minimal adversarial cost (*MAC*): The smallest adversarial cost A that can be obtained for instances in the negative class $\mathcal{X}_f^-$ of a deterministic classifier f. See 204.

Minkowski metric: A distance metric for the convex set $\mathbb{C}$ that is defined relative to a point $\mathbf{x}^{(c)}$ in the interior of the set. See 210.

near-isotropic: A set or body that is nearly round as defined by Equation 8.12. See 222–224, 226.

near-optimal evasion problem: A framework for measuring the difficulty for an adversary to find blind spots in a classifier using a probing attack with few queries. A family of classifiers is considered difficult to evade if there is no efficient query-based algorithm for finding near-optimal instances; see Chapter 8. See 11, 48, 200, 205.

negative class: The set of data points that are classified as negative by the classifier f (denoted by $\mathcal{X}_f^-$). See 26, 210, 211, 225, 231, 232.

norm ($\|\cdot\|$): A non-negative function on a vector space $\mathcal{X}$ that is zero only for the zero vector $\mathbf{0} \in \mathcal{X}$, is positive homogeneous, and obeys the triangle inequality. See 257.

normal instance: A data point that represents normal (allowable) activity such as a regular email message. See 26, *see also* data point.

obfuscation: Any method used by adversaries (particularly spammers) to conceal their malfeasance. See 8, 11, 40–42, 46, 112, 118.

https://doi.org/10.1017/9781107338548.014 Published online by Cambridge University Press

Ockham's Razor: An assumption that the simplest hypothesis is probably the correct one. See 22.

OD flow volume anomaly: An unusual traffic pattern in an OD flow between two points-of-presence (PoPs) in a communication network; e.g., a DoS attack. See 138.

one-class support vector machine: A formulation of the support vector machine used for anomaly detection. See *see also* support vector machine.

one-shot game: In game theory, any game in which players each make only a single move. See 58.

online learning: A learning process in which data points from the training dataset arrive sequentially. Often, online learning consists of sequential prediction followed by retraining as described in Section 3.6. See 25, 59–61, 235, 249, 252.

overfitting: A phenomenon in which a learned hypothesis fails to generalize to test data; i.e., it poorly predicts new data items drawn from the same distribution. Typically this occurs because the model has too much complexity for its training data and captures random fluctuations in it rather than the underlying relationships. Note, this phenomenon is distinct from nonstationarity; e.g., distributional shift. See 27.

PCA evasion problem: A problem discussed in Chapter 6 in which the attacker attempts to send DoS attacks that evade detection by a PCA subspace-based detector as proposed by Lakhina, Crovella, and Diot (2004*b*). See 142, 143.

performance measure: A function used to assess the predictions made by or actions taken by a learning agent. See 26, *see also* loss function.

polymorphic blending attack: Attacks proposed by Fogla & Lee (2006) that use encryption techniques to make intrusion instances indistinguishable from normal instances. See 41.

positive class ($\mathcal{X}_f^+$): The set of data points that are classified as positive by the classifier f (denoted by $\mathcal{X}_f^+$). See 26, 203, 210, 211, 225, 231.

positive homogeneous function: Any function p on a vector space $\mathcal{X}$ that satisfies $p(a\mathbf{x}) = |a|\, p(\mathbf{x})$ for all $a \in \Re$ and $\mathbf{x} \in \mathcal{X}$. See 214.

positive semi-definite function: A real-valued bivariate function, $k(\cdot\,,\,\cdot)$, on the space $\mathcal{X} \times \mathcal{X}$ is positive semi-definite if and only if $k(\mathbf{x}, \mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathcal{X}$. See *see also* kernel function.

prediction: The task of predicting an unobserved quantity about the state of a system based on observable information about the system's state and past experience. See 25.

prior distribution: A distribution on the parameters of a model that reflects information or assumptions about the model formed before obtaining empirical data about it. See 24, 107–110, 112, 278–280.

https://doi.org/10.1017/9781107338548.014 Published online by Cambridge University Press

probably approximately correct: A learning framework introduced by Valiant (1984) in which the goal of the learner is to select a hypothesis that achieve a low training error with high probability. See 15, 51, 55, 58, 65, 174, 176, 202.

probing attack: An attack that uses queries to discern hidden information about a system that could expose its weaknesses. See 11, 31, 40, 45, 61, 208, 227, 234, 251, *see also* near-optimal evasion problem.

query: A question posed to an oracle; in an adversarial learning setting, queries can be used to infer hidden information about a learning agent. See 10, 11, 18, 40, 42, 48, 49, 200, 201, 204, 205, 207, 209, 211, 212, 214–217, 220, 221, 225, 228, 229, 231, 232, 235–237, 244, 245, 264.

regret: The difference in loss incurred by a composite predictor and the loss of an expert used by the composite in forming its predictions. See 58, 60.

▶ **cumulative regret** ($R^{(m)}$): The total regret received for the m^{th} expert over the course of K rounds of an iterated game. See 60.

▶ **instantaneous regret** ($r^{(k,m)}$): The difference in loss between the composite predictor and the m^{th} expert in the k^{th} round of the game. See 59.

▶ **worst-case regret** (R^{*}): The maximum cumulative regret for a set of M experts. See 60.

regret minimization procedure: A learning paradigm in which the learner dynamically reweighs advice from a set of experts based on their past performance so that the resulting combined predictor has a small worst-case regret; i.e., it predicts almost as well as the best expert in hindsight. See 60.

regularization: The process of providing additional information or constraints in a learning problem to solve an ill-posed problem or to prevent overfitting, typically by penalizing hypothesis complexity or introducing a prior distribution. Regularization techniques include smoothness constraints, bounds on the norm of the hypothesis, $\|f\|$, and prior distributions on parameters. See 27.

reject on negative impact: A defense against *Causative* attacks, which measures the empirical effect that each training instance has when training a classifier with it, identifies all instances that had a substantial negative impact on that classifier's accuracy, and removes the offending instances from the training set, $\mathbb{D}^{(\text{train})}$, before training the final classifier. See 55, 114, 120, 128, 129, 131, 243.

residual: The error in an observation that is not accounted for by a model. Models such as PCA select a model according to the total size of their residuals for a given dataset. See 136, 145, 148–150, 152, 153, 159, 160, 163.

▶ **residual rate**: A statistic that measures the change in, the size of the residual caused by adding a single unit of traffic volume into the network along a particular

https://doi.org/10.1017/9781107338548.014 Published online by Cambridge University Press

OD flow. Alternatively, it can be thought of as a measure of how closely a subspace aligns with the flow's vector. See 152–154.

▶ **residual subspace**: In subspace estimation, the residual subspace (or abnormal subspace) is the orthogonal complement to the normal subspace used by the model to describe the observed data; i.e., the error component of each data point lies in the residual subspace. See 136, 139, 152.

response space ($\mathcal{Y}$): The space of values for the response variables; in classification this is a finite set of categories and in binary classification it is {"+", "−"}. See 23, 26.

response variable: An unobserved quantity that is to be predicted based on observable explanatory variables. See 23, *see also* label.

risk ($R\left(P_{\mathcal{Z}}, f\right)$): The expected loss of a decision procedure f with respect to data drawn from the distribution $P_{\mathcal{Z}}$. See 26, 171, 176, 177, 180, 185, 188.

robust statistics: The study and design of statistical procedures that are resilient to small deviations from the assumed underlying statistical model; e.g., outliers. See 55.

RONI: See vii, 120, 128–132, 243, *Glossary:* reject on negative impact.

scale invariant: A property that does not change when the space is scaled by a constant factor. See 208.

secure learning: The ability of a learning agent to achieve its security goals in spite of the presence of an adversary that tries to prevent it from doing so. See 7.

security goal: Any objective that a system needs to achieve to ensure the security of the system and/or its users. See 31.

security-sensitive domain: A task or problem domain in which malicious entities have a motivation and a means to disrupt the normal operation of system. In the context of adversarial learning, these are problems where an adversary wants to mislead or evade a learning algorithm. See 3, 5, 6, 29, 31.

set indicator function: The function $\mathrm{I}_{\mathbb{X}}\left[\,\cdot\,\right]$ associated with the set $\mathbb{X}$ that is 1 for any $x \in \mathbb{X}$ and is 0 otherwise. See 256.

shift invariant: A property that does not change when the space is shifted by a constant amount. See 208.

stationarity: A stochastic process in which a sequence of observations are all drawn from the *same* distribution. Also, in machine learning, it is often assumed that the training and evaluation data are both drawn from the same distribution—we refer to this as an assumption of stationarity. See 22, 31.

support vector machine: A family of (nonlinear) learning algorithms that find a maximally separating hyperplane in a high-dimensional space known as its reproducing kernel Hilbert space (RKHS). The kernel function allows the method to compute inner

https://doi.org/10.1017/9781107338548.014 Published online by Cambridge University Press

products in that space without explicitly mapping the data into the RKHS. See 39, 43, 46, 52, 55, 63, 171, 176.

SVM: See 55, *Glossary:* support vector machine.

threat model: A description of an adversary's incentives, capabilities, and limitations. See 13, 31.

training: The process of using a training dataset $\mathbb{D}^{(\text{train})}$ to choose a hypothesis f from among a hypothesis space, $\mathcal{F}$. See 25, 33, 34, 40, 44, 131, 134, 142–144.

training algorithm ($H^{(N)}$): An algorithm that selects a classifier to optimize a performance measure for a training dataset; also known as an estimating procedure or learning algorithm. See 24, 25, 32, 203, 234, 247.

training dataset ($\mathbb{D}^{(\text{train})}$): A dataset used by a training algorithm to construct or select a classifier. See 6, 17, 18, 21, 25, 27, 30–32, 34, 36, 39, 40, 45, 46, 49, 55, 56, 58, 61, 63, 69, 71, 72, 94, 106, 107, 114–116, 119, 120, 128, 134, 162, 171, 197, 247, 278, 281, *see also* dataset.

true-positive rate: The frequency for which a predictor correctly classifies positive instances. This is a common measure of a predictor's performance and is one minus the false-negative rate. See 151, *see also* false-negative rate.

unfavorable evaluation distribution: A distribution introduced by the adversary during the evaluation phase to defeat the learner's ability to make correct predictions; this is also referred to as *distributional drift*. See 45.

VC-dimension: The VC or Vapnik-Chervonenkis dimension is a measure of the complexity of a family of classifiers, which is defined as the cardinality of the largest set of data points that can be shattered by the classifiers. See 174, 176, 179, 205.

vector space: A set of objects (vectors) that can be added or multiplied by a scalar; i.e., the space is closed under vector addition and scalar multiplication operations that obey associativity, commutativity, and distributivity and has an additive and multiplicative identity as well as additive inverses. See 257, 259.

virus detection system: A detector tasked with identifying potential computer viruses. See 31.

https://doi.org/10.1017/9781107338548.014 Published online by Cambridge University Press