

## ta

lobrem ipsum doior sit amet, consēctetur adipiscing eint. Ut purus eint, vestibulum ut,  
 placere ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero,  
 nonummy eget, consēctetur eint, vulpulate a ma. Donec vestibula augue eu neque.  
 Consequat eint, morbi tunc, ut neque et neque et neque et neque ac tunc  
 egestas, mauris ut leo. Cras viverra mattis rhoncus semper. Nulla et lectus vestibulum urna  
 fringilla ultrices. Phasellus ut tellus sit amet tortor gravida placet. Integer sapien est,  
 iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum.  
 Aenean faucibus. Morbi dolor nulla, mauesuada eu. pulvinar at, mollis ac, nulla. Curabitur  
 auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan  
 eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.  
 Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem  
 justo. Nam lacus libero, pretium at, lobortis vitae, ultrices et, tellus. Donec aliquet, tortor  
 sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac  
 orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla.  
 Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus.

## Ranking & probabilities

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut,  
 Curabitur ut, felis. Curabitur ut, felis. Curabitur ut, felis. Curabitur ut, felis. Curabitur ut, felis.  
 Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis  
 egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna  
 fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est,  
 ac acen in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum.  
 Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur  
 auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan  
 sed, dui. Morbi ut tortor. Donec dui. Donec dui. Donec dui. Donec dui. Donec dui.  
 Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non  
 justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor  
 sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac

orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

## beyond binary classification

### multi classes

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

### regression

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

### unsupervised & descriptive learning

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

## Distance-based models

similarity is some function of distance. clustering is grouping data without prior information(unlabeled data). **Why cluster?** •make apparent the natural groupings/structure in the data (perhaps for further processing) •To discover previously unknown relationships •To provide generic labels for the data **Clustering** we organize data into classes such that: •The within-class (intra-class) similarity is high(Lower intra-class variance) •The between-class (inter-class) similarity is low(Higher inter-class variance) •Objects in the same group (a cluster) are more similar to one another than to objects in other groups (clusters)

**Distance measures** D(x1,x2) is a function D:  $X \times X \rightarrow R$  such that for any xyz in X: 1. D(x,x) = 0 2. If x != y then  $D(x, y) > 0$  3. D(x,y) = D(y,x) 4.  $D(x, z) \leq D(x, y) + D(y, z)$ . **Common Distance measures**Hamming(count differences), Manhattan, Euclidian, Minkowski( $L_p$ ), Chebyshev, Mahalanobis

**Distance-based methods** Methods for classification and clustering based on distances to exemplars or neighbors. Exemplar - a prototypical instance. Neighbor – a nearby instance or exemplar.

**1NN** Assign the new instance x to the nearest labeled training point (or exemplar).

•Training = memorizing the training data •Each point is an exemplar, or exemplars are computed from the data •But it generalizes, unlike the lookup table approach •The implicit decision boundaries of a 1NN classifier comprise a Voronoi tessellation, Leads to piecewise linear decision boundaries.

### Probabilistic models

skipped

### Features

**features (aka attributes)** A mapping from the instance space X to the feature domain F. model can be thought of as just a new feature, particular combination of the input features, constructed to solve the task at hand. **Types** Quantitative(numeric scale, age), Ordinal(ordered set, ranks), Categorical(unordered set, color). **Statistics** Mode, Median, mean, range, std, Skewness, Kurtosis, Percentiles, Deciles, Quartiles, Interquartile range **Plots** histogram, Percentile plot(CDF, Shows cumulative fraction of data - what % fall below a given value)

**Feature construction and selection** constructed new features from the given feature set, then select a suitable subset prior to learning(kernel methods, N-grams, Cartesian products of categorical features.)

**kNN** •Classify a new instance by taking a vote of the  $\geq 1$  nearest exemplars. •vote among all neighbors within a fixed radius r •combine the two, stopping when *count* > k or *dist.* > r •distance weighting, the closer an exemplar is to the instance, the more its vote counts. **ties** Preference to the 1NN, or Random choice. **Pros** train  $O(n)$ , simple, intuitive **Cons** test  $O(n)$  require a good deal of storage, and can't easily represent a specific boundary geometry, rely on a useful distance metric.

**Clustering vs. classification** In a classifier, possible class labels are provided. In a clustering problem, possible labels are the cluster labels learned from the training set. Assigning of labels or clusters to data points is classification.

**Clustering** find clusters (groupings) that are compact with respect to the distance metric. **Scatter Matrix** The scatter of X is defined as the trace of the scatter matrix.(The trace is the sum of the diagonal elements of a square matrix).

**Kmeans** NP-complete, no efficient solution to find the optimal clustering (data partition). **K-means algorithm** heuristic algorithm, not optimal, converge to local optimal, works quite well in most cases. run several times (with a random starting point) and then the best solution is selected(the solution with the smallest within-cluster scatter).

**Kmedoids** medoid of a set of points is the point with the minimal average dissimilarity (distance) to all other points in the set.

**PCA** dimensionality reduction, finding the intrinsic linear structure in the data. eigenvalues can give clues to the inherent dimensionality of the data.

### Model ensembles

are â€œmetaâ€œ methods. Construct different models from adapted versions of the training data, Combine the predictions by averaging voting, weighted voting. a strong classifier from

a set of weak classifiers. **Pros** improve performance, avoid over-fitting. **Bagging** bootstrap aggregation. T models on different samples. Samples are training data with replacement choosen with uniform probability, called bootstrap sample. Decision boundary is *piecewise linear*.(useful for tree model). **Subspace sampling** Build each decision tree from a different random subset of the features. *Bagging + subspace**sampling = randomforests**method*.

**Boosting** create diverse training sets, add classifiers that do better on the misclassifications from earlier classifiers, by giving them a higher weight(less susceptible to overfitting).Can be extended to multi-class classification. **Adaboost**(adaptive boosting), As long as the performance of each classifier is better than chance*c* < 0.5 , it is guaranteed to converge to a better classifier.**Procedural.** Train a classifier assign it a confidence factor based on the error rate 2. Give misclassified instances a higher weight. 3. Repeat for T classifiers 4. The ensemble predictor is a weighted average of the models (rather than majority vote) 5. Threshold for binary output.

**Boosting1 vs bagging2** •1 Can achieve zero training error by focusing on the misclassifications. A bias reduction technique (increase accuracy) •2 With relatively large bootstrap sample sets, there tends to be little diversity in the learned models. A variance-reduction technique (increase consistency).

### ML experiments

**ML experiments** confirm the various assumptions in a ML problem, establish realistic expectations for performance. **Questions** specific model perform on data from D? a set of models has the best performance on D? models from learning algorithm A perform on D? learning model A produces the best models for domain D?

**Performance measure** accuracy assume the same class distribution of read and test. average recall assume real problem is uniform class distribution.

**Performance estimation**(testing). Holdout Method(leave ~ 30%), cross-validation, leave-one-out.

**Cross-validation** reduce the sample variance by  $1/\sqrt{k}$ . large data sets, fewer folds. For sparse, leave-one-out best. If *var* > *acceptable var*, we need more data. If statisfied, run over the entire data set produce the final model.

**Testing** Applied after the model is finalized.

### Neural network

**When to use** input high dim. target function unknown. long training time. human readability. good for complex pattern recognition.

**RNN** Directed cycles exist.

**GD** search *hypothesis* space to find *w* for minimum error(slow, local minimum). *Variation* incremental GD, SGD.

**BP** minimize the squared error, iteratively propagate errors backwards from output units. $\Delta w_{ji} = \eta \delta_j z_{ji}$ . **inductive bias** smooth interpolation between data points.

**termination condition** number of iterations, predetermined error threshold. **Overfitting** every iteration measure error on the validation set(cross-validation approach).

**RL** agent learns behavior through trial-and-error interactions with a dynamic environment based on a reward signal(Game playing). **Rewards** from sequences of actions, can be separated temporally. **Q-learning** finds optimal action-selection policy for Markov Decision Process, learns an action-value function that gives the expected utility(convergence slow).

**Multi-label classification** class labels are not mutually exclusive.feature-to-class mapping, dependence between classes are learned.(blog post tags)

**Transfer learning**(inductive transfer) apply knowledge to a different but related problem(walk to run)

**Online (incremental) learning** data available sequentially over time. updates the model each time a new data point arrives(visual tracking)

**Active learning** semi-supervised learning, the algorithm queries information source(the user) to obtain the desired model(robotics, where to point the camera or sensors to gain useful information)

**DL** learn many-layered NN, more abstract features. learn good representations of the data through unsupervised learning. **Cons** complex, slow, hard to explain.