•Validation: If the test set is used to select or refine models, it is called validation (or development) (test) set• Test: Estimate accuracy of the model• Model Deployment: If the accuracy is acceptable, use the model to classify new data

**Decision tree** **Pro’s** • Easy to explain (even for non-expert) • Easy to implement (many software)• Efficient• Can tolerant missing data• White box• No need to normalize data• Non-parametric: No assumption on data distribution, no assumption on attribute independency• Can work on various attribute types

• **Con’s**• Unstable. Sensitive to noise• Accuracy may be not good enough (depending on your data)• The optimal splitting is NP. Greedy algorithms are used

Overfitting **Info gain**: ID3/ C4.5 **Higher entropy -> higher uncertainty**

• What is the range of Gini index?• The minimum= 0, meaning pure• The maximum=? What is the case that Gini index reach the maximum?

• Answer: when is n split, (1/n, 1/n, …. 1/n), then 1 – (1/n\*1/n)\*n = 1- 1/n

• The **weakness** of three measures: • Information gain: • biased towards multivalued attributes

• Gain ratio: • tends to prefer unbalanced splits in which one partition is much smaller than the others

• Gini index: • biased to multivalued attributes• has difficulty when # of classes is large • tends to favor tests that result in equal-sized partitions and purity in both partitions

• 避免决策树过拟合：1交叉验证 2 预先剪枝 3 后剪枝

Bayes' Theorem Example 1: Cancer Tests 
Test Pos 
Test Neg 
Test Pos 
Test Neg 
Cancer (1%) 
20% 
Cancer (1%) 
True Pos 
1% x = .008 
False Neg 
1% x = .002 
No Cancer (99%) 
9.6% 
90.4% 
No Cancer (99%) 
False Pos 
99% x 9.6% = .09504 
True Neg 
99% x 90.4% = .89496 
C] Only 1% people have cancer 
How accurate is the test? 
80%? 99%? 
P(XH) = 
C] Chance of true positive is thus 
= 0.008 
C] According to Bayes' Theorem, P(H I X) = P(XI the chance of having 
a cancer given positive test results is 
True pos/ (True pos + False pos) = 0.008 / (0.008+0.09504) = 7.76% 
C] The Theorem lets us correct for the skewness introduced by false positives 

Bayes’ Theorem allows us to tell back and forth between posterior and likelihood(e.g., P (Rain | Cloud) and P(Cloud | Rain)), tests and reality, which is the most important trick in Bayesian Inference

**Naïve bayes:• Strength** •:A naïve Bayesian classifier, has comparable performance with decision tree and selected neural network classifiers

• Each training example can incrementally increase/decrease the probability that a hypothesis is correct—prior knowledge can be combined with observed data• **Weakness**• Assume attributes are conditional independence, therefore loss of accuracy

• **Logistic regression**:• **Pros**• Can handle multiple types of features• Fast and easy• Generally speaking, more robust and better performance than tree• Interpretable: both weights and predicted value (• Predicted value: probability• Weights: effect of the feature. Unit change of log odds)

• **Cons**• Linear model: if the decision boundary is not linear, then LR is not good

• **Generative && discriminative classifiers**• X: observed variables (features) • Y: target variables (class labels)

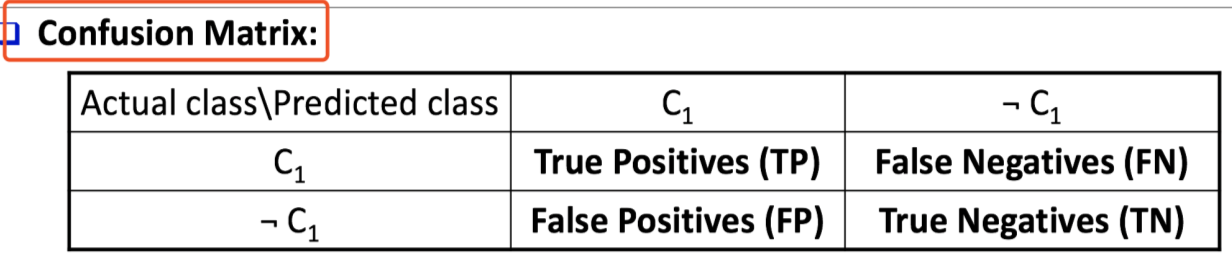
• A generative classifier models p(Y, X)(important)• It models how the data was "generated"? • "what is the likelihood this or that class generated this instance?" and pick the one with higher probability( • Naïve Bayes • Bayesian Networks )

• A discriminative classifier models p(Y|X) • It uses the data to create a decision boundary(• Logistic Regression• Support Vector Machines)

**discriminative classifier**• **Strength**• Prediction accuracy is generally high, As compared to generative models • Robust, works when training examples contain errors• Fast evaluation of the learned target function, Comparing to (covered in future) Bayesian networks (which are normally slow)

• **Weakness**• Long training time• Difficult to understand the learned function (weights): Bayesian networks can be used easily for pattern discovery

• Not easy to incorporate domain knowledge (P(X|Y), like what a patient under a disease would look like):Easy in the form of priors on the data or distributions



Accuracy = (TP + TN)/All Error rate: 1 – accuracy Sensitivity = TP/P, true positive rate Specificity = TN/N Precision = TP/(TP +FP)Recall=TP/(TP +FN)

Fmeasure = ((β^2+1)P∗R)/(β^2 P+R) F\_1=(2P∗R)/(P+R)

ROC: Vertical axis represents the true positive rate (TP/P) Horizontal axis rep. the false positive rate (FP/N) AUC: Area Under Curve is a measure of the accuracy of the model

MODEL SELECTION ISSUE: Accuracy, Speed, time to construct and use model, Robustness, Scalability: efficiency in disk-resident databases, Interpretability

如何把多分类问题转换为二分类？1 one vs rest. 取出一个类，把其他类统一看成反类。2 one vs one. 类两两之间训练一个分类器。

**• Random forest and XGBoost** are the most commonly used algorithms for tabular data:

• **Pros**• Good performance for tabular data, requires no data scaling• Can scale to large datasets• Can handle missing data to some extent

• **Cons**• Can overfit to training data if not tuned properly•Lack of interpretability (compared to decision trees)

Ensemble method have two high level idea: bagging, boosting. If decision tree is the base for bagging, then it is random forest.

Boosting has two famous implementation: adaboost, gradient boosting. XGboost is the scalable version of gradient boost.

**Bayesian belief network**: Allows class conditional independencies between subsets of variables

Two components: A directed acyclic graph (called a structure) , A set of conditional probability tables (CPTs)

SVM:

We allow data points to be on the “wrong side” of the margin boundary Penalize points on the wrong side according to its distance to the margin boundary

𝜉 : slack variable C (> 0): Controls the trade-off between the penalty and the margin Smaller C: allow more mistake

Kernel function: mapping low dimen data to high dimen data so that separable.

• SVM is effective on high dimensional data

• The complexity of trained classifier is characterized by the # of support vectors rather than the dimensionality of the data

• The support vectors are the essential or critical training examples—they lie closest to the decision boundary(hyperplane) (MMH)

• Thus, an SVM with a small number of support vectors can have good generalization, even when the dimensionality of the data is high

• SVM is not scalable to the # of data objects in terms of training time and memory usage

• Features: training can be slow but accuracy is high owing to their ability to model complex nonlinear decision boundaries (margin maximization)

• Used for: classification and numeric prediction• SVM can also be used for multiple classify and regression analysis (with additional parameters)

• Applications: • handwritten digit recognition, object recognition, speaker identification, benchmarking time-series prediction tests

• **SVM** • **Pros**• Elegant mathematical formulation, guaranteed global optimal with optimization• Trains well on small data sets• Flexibility through kernel functions• Conformity with semi-supervised training• **Cons**• Not naturally scalable to large data sets

• **Deep learning** • **Pros**• Very good performance on certain tasks, for certain types of data• Requires very little feature engineering

• Good generalization • E.g. models trained on ImageNet dataset for classification can help tasks such as segmentation

• **Cons**• Requires huge amounts of computation power• Black box model• Hard to tune the architecture and hyperparameters for new tasks.

Why CNN not MLP: computational expensive and hard to train. CNN: para sharing, equivariance, subsample, RNN: feedback loop, longterm dependency

• **Pattern based methods** (59 - 66)• **Pros**: explainable, white box • **Cons**: accuracy is low

Rule Conflict solution: size order, class-based order, rule-based order **Pattern-based classification**: An integration of classify and frequent pattern mining

Why pattern-based classification? • Feature construction. Higher order; compact; discriminative• Complex data modeling, Graphs, Sequences

**KNN**

• The number of neighbors k. • Small k: overfitting (high var., low bias). • Big k: bringing too many irrelevant points (high bias, low var.)

**Pros**: easy to implement, no time for train. Robust to noisy data: local consideration **Cons**: Curse of dimension: distance between neighbors could be dominated by irrelevant attributes

K-Means++: The first centroid is selected at random. The next centroid selected is **the one that is farthest from the currently** selected

K-Medoids: **medoids** can be used, which is **the most centrally located object in a cluster**

K-Medians: medians are used (**L1-norm** as the distance measure)

K-Modes: replacing means of clusters with modes and can handle **categorical** data

**AGENES**

Single link (nearest neighbor)

Complete link (diameter, the farest neighbor)

Average link 两个簇里面任意一对点的距离的平均

Centroid link (centroid similarity, the distance between two center)

In exam, show dendrogram of clustering. And steps of calculation.

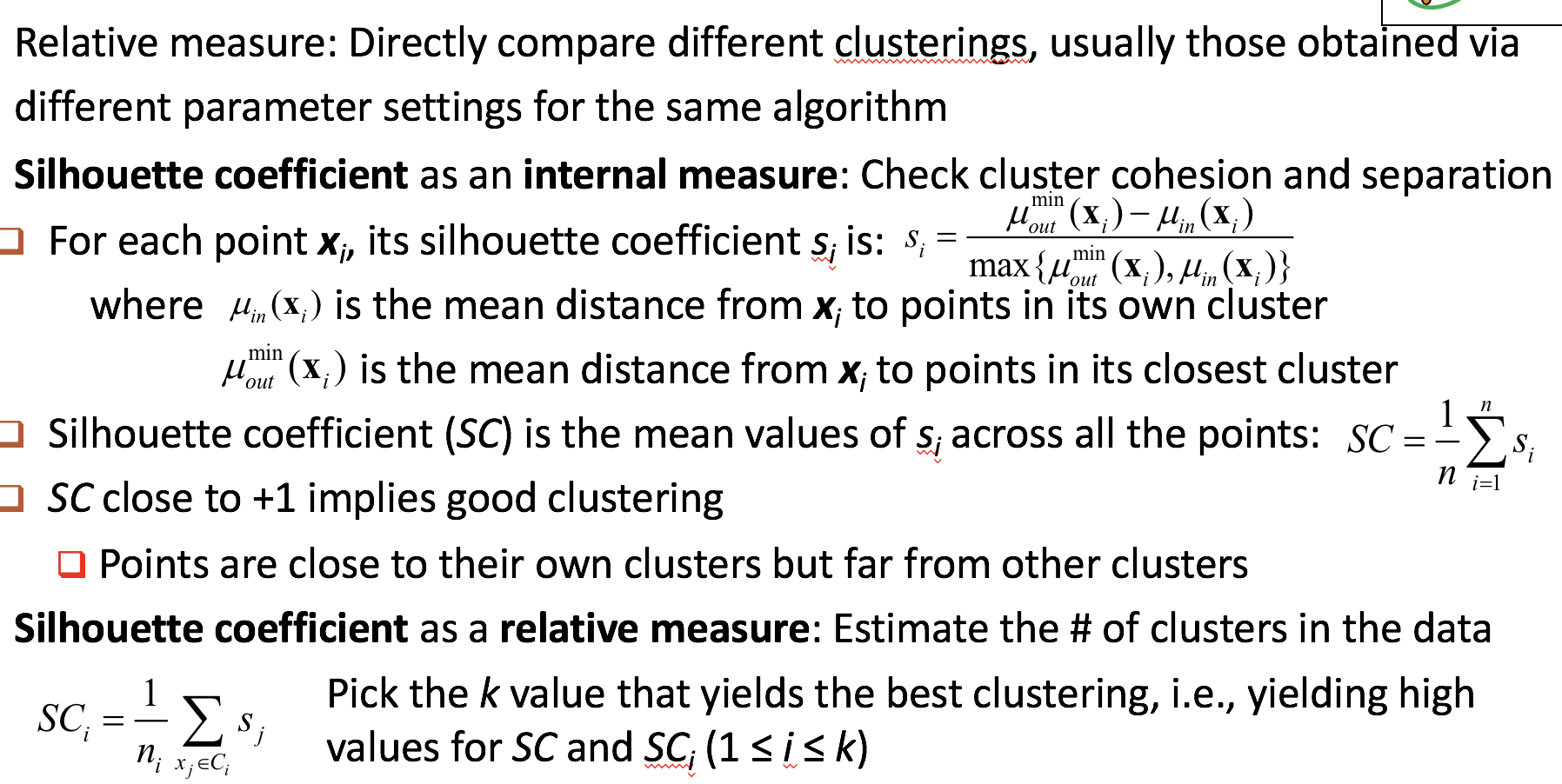
**Hierarchical clustering weakness**:• Can never undo what was done previously• Do not scale well. Time complexity O(n2)

DBSCAN: O(nlogn), O(n2). Sensitive to parameter choice.

**Internal measure**: Unsupervised, criteria derived from data itself. Evaluate the goodness of a clustering by considering how well the clusters are separated and how compact the clusters are, e.g., silhouette coefficient

**Relative measure**: Directly compare different clustering, usually those obtained via different parameter settings for the same algorithm

**Silhouette coefficient**. Very important!!



**Gaussian mixture model:**

**Pros**: Mixture models are more general than partitioning: different densities and sizes of clusters. Clusters can be characterized by a small number of parameters

The results satisfy the statistical assumptions of generative models

**cons**: Converge to local optimal. Computationally more expensive. Hard to estimate the number of clusters. Can only deal with spherical clusters

CLIQUE: density-based and grid-based subspace clustering

Right-most path extension: convert graph to sequence

Right-most path: The path from root to the right-most leaf (choose the vertex with the smallest index at each step) small-leaf first search

Closed pattern: lossless, no super pattern with same support

Max pattern: lossy, no super pattern frequent