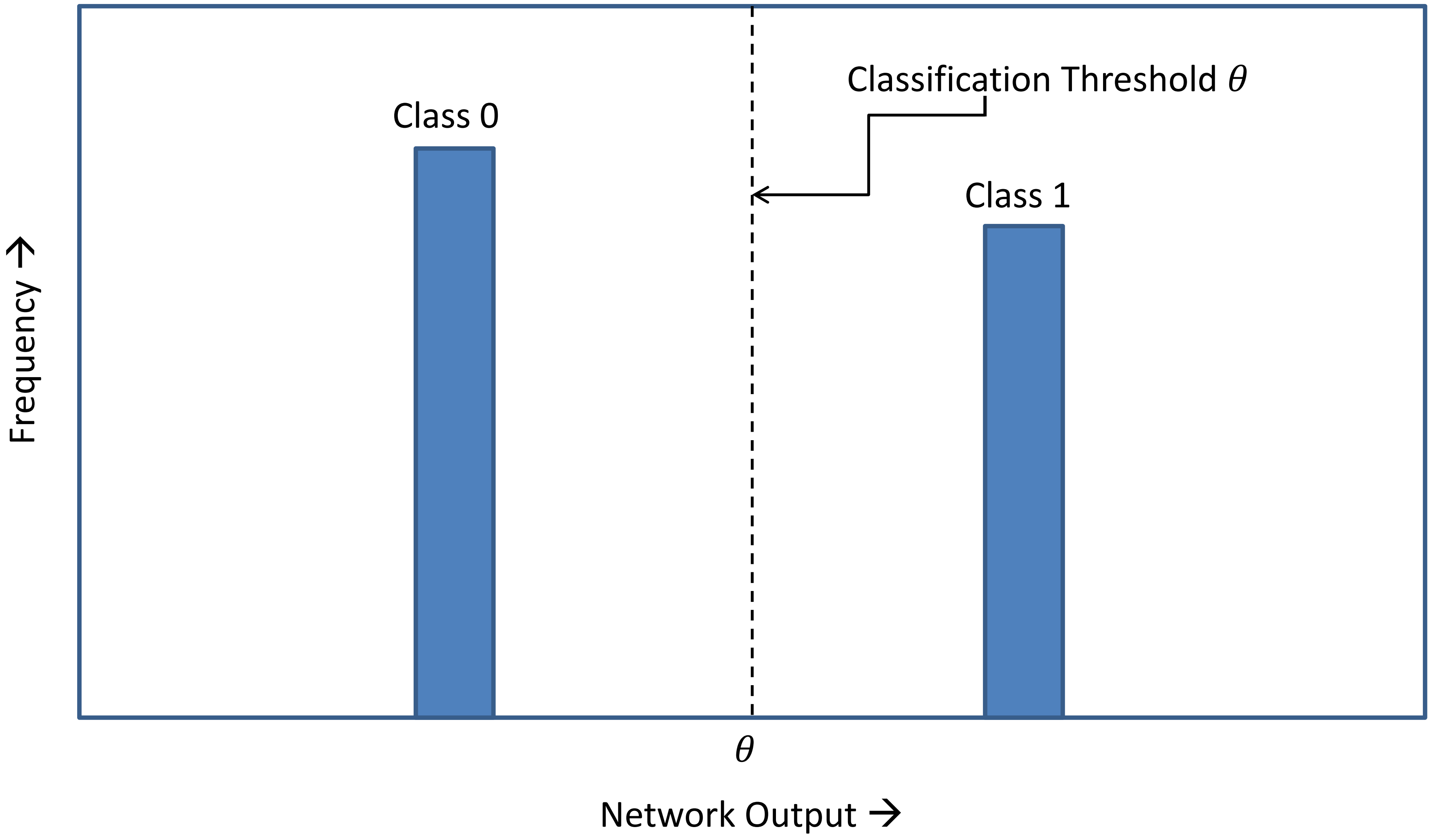
Machine Learning
Baruch College
Lecture 4

Miguel A. Castro

Today:

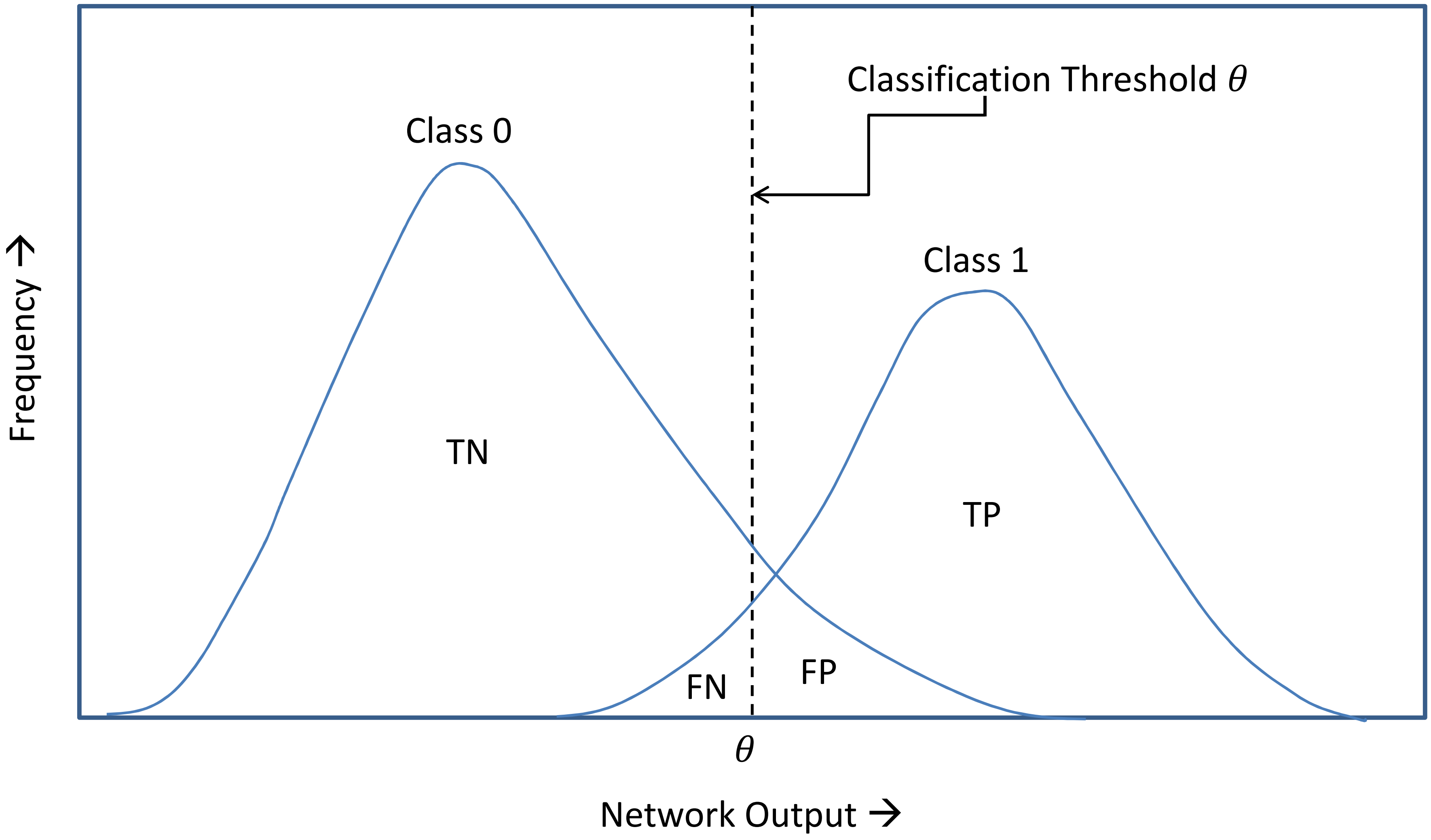
- Announcements:
  - – Remember your *Project Proposal*: Due one week from today, Wednesday April 29.
  - – *Exam* will be two weeks from today, Wednesday May 6.
- Review of Last Lecture.
- Association Rules.
- Brief intro to Time Series Forecasting with ANNs
- Talk about any project ideas you may have.

Last Time… Ideal Classifier



- The ideal classifier is able to separate the classes easily (remember the AND, OR, XOR examples earlier).

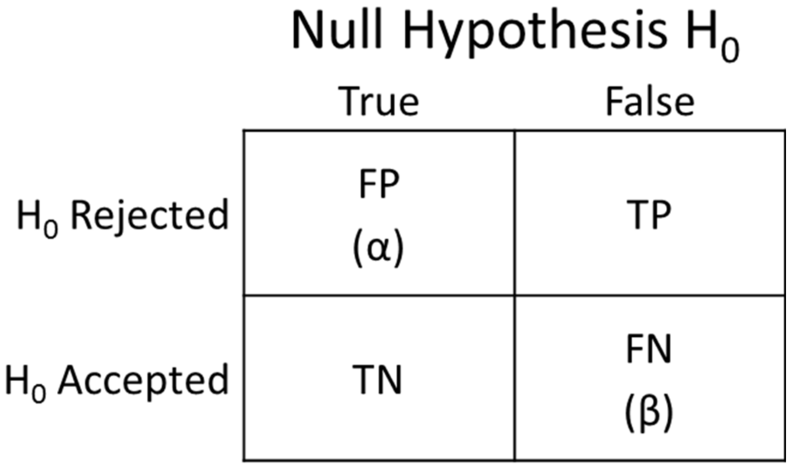
Last Time… Noisy Classifier



- In the presence of noise, classes are blurred and there is an area of confusion (overlap) no matter where θ is placed. Notice tradeoff between FN and FP as θ is moved.

Some Classifier Vocabulary

- *ROC Curves* used to study Classifier Performance
- *Confusion Matrices* also used to study Classifier Performance.
- Can couch in the language of *Hypothesis Testing*:



	Null Hypothesis H_0	
	True	False
H_0 Rejected	FP (α)	TP
H_0 Accepted	TN	FN (β)

- α = Prob of Type I Error (FP); $1 - \alpha$ = “Confidence”
- β = Prob of Type II Error (FN); $1 - \beta$ = “Power”

Classifier Performance

- To compare classifiers, compare their ROC curves.
- Ranking Criterion: the larger the Area Under the Curve, *AUC*, the better the classifier.
- Other comparison metrics: *Youden Index*, *Euclidean Distance Accuracy*, *Chi-Squared Metric* comparing against “Random Classifier,” *Yule’s Q*, …
- These do not take Misclassification Costs into account.

Costs and ROC Analysis

- Suppose a trained classifier neural network's output, o, represents the probability that Class 1 is True. (**Aside*: an interesting Class Project would be to find out how a Classifier Neural Network with logistic output activation compares to Logistic Regression.)
- For example, Class 1 could mean "A Trade is *Profitable*," while Class 0 could represent "A Trade is *Unprofitable*."
- Generally, there are different costs/benefits associated with the different outcomes.
- Let P be the Profit associated with entering a Profitable Trade, and L be the Loss associated with entering an Unprofitable Trade.
- We should enter a Trade only when the Expected Profit exceeds the Expected Loss:

$$oP > (1 - o)L.$$

Costs and ROC Analysis

- Since our classification of a Class 1 (i.e. a Profitable Trade) is made whenever the network's output o exceeds the threshold θ:

$$o > \theta,$$

- This immediately suggests an expression for the threshold for entering profitable trades, independent of classifier performance:

$$\theta = \frac{L}{L + P}.$$

- In HFT usually $L > P$ due to transaction costs and other frictions.

Incorporating Classifier Performance: Misclassification Costs

- What about *Misclassification Errors*?
- We can see that Profit will only result from entering a Profitable Trade:

$$Expected\ Profit = TP \cdot P.$$

- A Loss can occur either from entering an Unprofitable Trade, or from *not* entering a potentially Profitable Trade:

$$Expected\ Loss = FP \cdot L + FN \cdot P.$$

Incorporating Classifier Performance: Misclassification Costs

- The condition for profitability is that the Expected Profit must exceed the Expected Loss:

$$TP \cdot P > FP \cdot L + FN \cdot P, or:$$

$$(TP - FN) \cdot P > FP \cdot L.$$

- This imposes a condition on the classifier's performance (and misclassification costs) given the trade's profit and loss:

$$\frac{(TP - FN)}{FP} > \frac{L}{P}.$$

- The left-hand side can be thought of as a ***performance criterion*** for any classifier used for similar purposes.

Misclassification Costs, More Generally

- Generally, there is a tradeoff between Type I Error (FNs) and Type II Error (FPs).
- This can be incorporated into performance metrics to adjust for misclassification costs.
- For example, we can adjust the Euclidean Distance Accuracy to introduce a compromise between FPs and FNs as follows:

$$Cost\ Adjusted\ Dist\ Acc = 1 - \sqrt{W(1-tp)^2 + (1-W)fp^2},$$

  where $0 < W < 1$, and where W penalizes FNs, and the FPs are penalized by $(1 - W)$.

- This cost adjustment models the tradeoff between FPs and FNs, and can be used in the other performance metrics as well.
- Note than in biological organisms, W is generally larger than $1 - W$, *i.e.*, $W > 1/2$ (Type II Errors are generally more costly).
- This is not always the case, as we saw in our trading example, where FPs were costlier than FNs due to transaction costs. This also holds in a court of law (convicting an innocent person is more detrimental), etc.

Association Rules

- *Association Rules* (also known as Market Basket Analysis) are an example of a *Nonparametric* Technique, and of *Unsupervised Learning*.
- Association Rules are of the form "If X then Y."
- X is the "*Antecedent*" and Y is the "*Consequent*."
- But this distinction is *artificial* when dealing with *temporal concurrency*; however, we can introduce a *time delay* between Antecedent and Consequent, which can lead to prediction.
- Example: A department store collects information on customer transactions. We wish to find Association Rules of the form: *"If jeans and t-shirts are purchased together, then belts are also purchased often."*
- Or, *"If a customer is female and she purchased vitamin supplements then calcium supplements are often also purchased."*
- It is unclear which is the Antecedent and which the Consequent.
- The idea is to discover these association rules *Automatically.*
- We wish to extract rules that are *Actionable*.
- Problem: The Number of possible items (occurrences) can be very large, and number of transactions can be huge.

Last Time: Association Rules

- In HFT, construct Consequents by introducing <u>*time delay*</u> (*e.g.*, search for price movements that exceed costs), then find relevant Antecedents (as a way to prune the search space).
- E.g.: X = "*INTC price went down and OIH went up this second.*" Y = "*One second later, AMAT goes down.*"
- Still cannot infer causality (there could be a common cause, etc.).
- In practice, we often care about the repeatability of the temporal pattern regardless of the true causal connection or "story".
- Hypothesis: "*AMAT will go up by more than transaction costs in the next second (or relevant time period).*"
- Actionable: "*Buy AMAT now.*"

Association Rules

- More formally, Let $\mathcal{I} = \{i_1, i_2, \dots, i_n\}$ be a set of literals called *Generalized Items*. These could be all the items sold at a store, along with customer demographic information, item category information, etc., or they could be the S&P500 stock movements along with industry categories, macroeconomic states, etc.
- Let $\mathcal{D}$ be a *Database* of m transactions (or occurrences) within a specified period of interest: $\mathcal{D} = \{d_1, d_2, \dots, d_m\}$ is a set of m binary n-tuples $\{0,1\}^n$ where 0/1 represents the presence/absence or occurrence/non-occurrence of a Generalized Item.
- We call a subset X of $\mathcal{I}$ an *Itemset*.

The Database of Transactions

$$\mathcal{D}$$

$$\begin{array}{c|ccccccc|}
 & i_1 & i_2 & i_3 & i_4 & \cdots & i_{n-1} & i_n \\
\hline
d_1 & 0 & 1 & 1 & 0 & \cdots & 0 & 1 \\
d_2 & 0 & 0 & 1 & 0 & \cdots & 1 & 0 \\
d_3 & 1 & 0 & 1 & 0 & \cdots & 0 & 0 \\
d_4 & 0 & 0 & 0 & 0 & \cdots & 1 & 0 \\
\vdots & & & & \vdots & & & \\
d_{m-1} & 1 & 1 & 0 & 1 & \cdots & 0 & 1 \\
d_m & 0 & 0 & 1 & 0 & \cdots & 0 & 0 \\
\hline
\end{array}$$

Association Rules

- Suppose we have two disjoint itemsets $X \subset \mathcal{J}$ and $\mathrm{Y} \subset \mathcal{J}$ such that $X \cap Y = \emptyset$
- We say there is an *<u>Association Rule</u>* "$X \Rightarrow Y$" if both itemsets are frequently present together in the same transaction (or basket, or occurrence).
- For example, if $X = \{i_2, i_7\}$ and $Y = \{i_3\}$, the Rule $X \Rightarrow Y$ can be interpreted as saying: "*When items i_2 and i_7 are present in the same transaction (occur together), then item i_3 is often also present (often also occurs).*"
- Notice that the requirement that $X \cap Y = \emptyset$ ensures that we don't end up with trivial associations like if i_2 and i_3 are present (occur) then i_3 is also present (occurs).

The Association Rule Problem

- Statement of the *Association Rule Problem*: "*Find Interesting Association Rules from the Database of Transactions $\mathcal{D}$.*"
- The key here is to define what "interesting" means, and to figure out an *automated* way that this computationally intense problem can be tractably solved.
- The approach most often used today (as found in the literature) is the *Support-Confidence Framework*.

The Support-Confidence Framework

- The Support-Confidence Framework defines an Interesting Association as follows:
- An Interesting Association Rule $X \Rightarrow Y$ occurs when:
  1. X and Y have support s, and
  2. X and Y have confidence c.
- *Support* is defined as a lower bound on the percentage of transactions in $\mathcal{D}$ that contain both itemsets X and Y; *i.e.*, $P(X \cap Y) \geq s$.
- *Confidence* is defined as the lower bound on the percentage of those transactions containing X that also contain Y; *i.e.*, $P(Y|X) \geq c$.

The Support-Confidence Framework

- For example, suppose the transaction database $\mathcal{D}$ contains 1 million transactions and 10,000 of those contain both itemsets X and Y.
- In this case, the support of $X \Rightarrow Y$ is

$$s = \frac{10^4}{10^6} = 1\%.$$

- Likewise, if 50,000 transactions contain X and, out of those, 10,000 also contain Y, then $X \Rightarrow Y$ has a confidence of

$$c = \frac{10^4}{5 \times 10^4} = 20\%.$$

Tractability (or lack thereof…)

- It is impossible to check all possible Itemset combinations for Interesting Association Rules.
- In fact, there are $\sum_{i=1}^{n} \binom{n}{i} = 2^n - 1$ such combinations, where n is the number of items, which could number in the thousands. For example, if $n = 1{,}000$ there would be over 10^{300} possibilities to check, which is prohibitive.
- However, we don’t have to check all possibilities…

Tractability

- Suppose we have found an Itemset that is supported. Then we know that all its subsets are supported (why?). For example, if $\{i_3, i_4, i_8\}$ is supported, then we know that $\{i_3\}$, $\{i_4\}$, $\{i_8\}$, $\{i_3, i_4\}$, $\{i_3, i_8\}$, etc… are all supported and we don’t need to check them.
- Also, all supersets of an Itemset that is not supported are themselves not supported (why?), so we don’t need to check them.
- Likewise, all supersets of a confident Itemset are confident, and all subsets of a non-confident Itemset are not confident. (why?)
- We can use these findings to prune the search space of Interesting Association Rules dramatically. We can start with 2-Itemsets and prune all supersets of those that are not supported, then explore 3-Itemsets etc. Then, we can prune all subsets of k-Itemsets that are not confident.

Tractability

- Another approach is to find viable antecedents to interesting consequents.
- Works well in HFT because interesting consequents are uncommon.
- Can be used with the other pruning techniques.
- Drawback: you may be missing interesting consequents (this is the whole point in ML: to generate useful information *automatically*; *i.e.*, you want to be “surprised” with interesting consequents that you didn’t know about *a priori*).

Throwing the Baby With the Bathwater

- I maintain that Support-Confidence does not necessarily Produce Useful Rules.
  1. The reason is that we should define Association Rules to be Interesting only when they deviate from Random Chance; *i.e.*, we want *<u>Non-Random Associations</u>*, which the Support-Confidence Framework does ***not*** distinguish from Random ones.
  2. Suppose Milk occurs in 60% of all transactions at a grocery store, and Bread occurs in 50% of all transactions. This means that *<u>By Random Chance Alone</u>* we would expect Milk and Bread to occur together in 30% of all transactions. This might seem like a high confidence number, so ***it would be selected as an interesting rule*** by the Support-Confidence Framework. But it ***means nothing***. It would be more interesting if they occurred together either significantly more frequently or significantly less frequently than the 30% expected by random chance alone.
  3. Support-Confidence ***can ignore anti-correlated occurrences***. For example, Coke and Pepsi may each occur in 50% of all baskets independently, (meaning we would expect them to occur together in 25% of all transactions by random chance alone). However, if they occur in 0.01% of all transactions together, this means they have a non-random effect on each other (*e.g.*: competition for market share?). Yet this might seem like a low support number, so we would be throwing away this information.

Throwing the Baby With the Bathwater (Continued)

4. Suppose SPY goes up in 50% of all short-period measurements and OIH goes up in 40% of all such measurements. This means that *By Random Chance Alone* we would expect SPY and OIH to go up together in 20% of all measurements. This might seem like a high confidence number, so ***it would be selected as an interesting rule*** by the Support-Confidence Framework. But it ***means nothing***. It would be more interesting if they occurred together either significantly more frequently or significantly less frequently than the 20% expected by random chance alone.
5. Support-Confidence ***can ignore anti-correlated occurrences***. For example, GLD and XLF may each go up in 50% of all measurements independently, (meaning we would expect them to go up together in 25% of all transactions by random chance alone). However, if they go up together in 1% of all transactions, this may mean they have a non-random effect on each other. Yet this might seem like a low support number, so we would be throwing away this information.
6. Let’s fix this...

The Correct Problem: Dependency

- Two Itemsets X and Y are (statistically) *Independent* if and only if the probability of their joint occurrence is just the product of their individual probabilities:

$$P(XY) = P(X)P(Y).$$

- Unlike in the Support-Confidence Framework, we are interested in Itemsets that are statistically *Dependent*:

$$|P(XY) - P(X)P(Y)| > \delta.$$

- In other words, we want their co-occurrence to *deviate statistically significantly* from what we would expect by random chance alone.
- The question is: what value should be chosen for δ?

The Correct Problem: Dependency

- An Association Rule $X \Rightarrow Y$ is called a *Dependence Rule* if X and Y are statistically Dependent.
- The correct statement of the problem is then: *"Find Interesting Dependence Rules from the Database of Transactions $\mathcal{D}$."*
- Or equivalently: *"Find Interesting Dependent Itemsets from the Database of Transactions $\mathcal{D}$."*
- We still need to define what we mean by "Interesting."

The Correct Problem: Dependency

- To test for the statistical significance of whether two Itemsets are Dependent, we can use the χ^2 Statistic:

$$\chi^2 = \sum_{\{minterms\}} \frac{(E(XY) - O(XY))^2}{E(XY)},$$

- Where $E(XY)$ is the *Expected* number of transactions where Itemsets X and Y would occur together if they were independent (the Null Hypothesis), and $O(XY)$ is the actually *Observed* number of transactions where X and Y occur together.
- Notice that the sum is over the *Minterms* of the Boolean product of X and Y…

The Correct Problem: Dependency

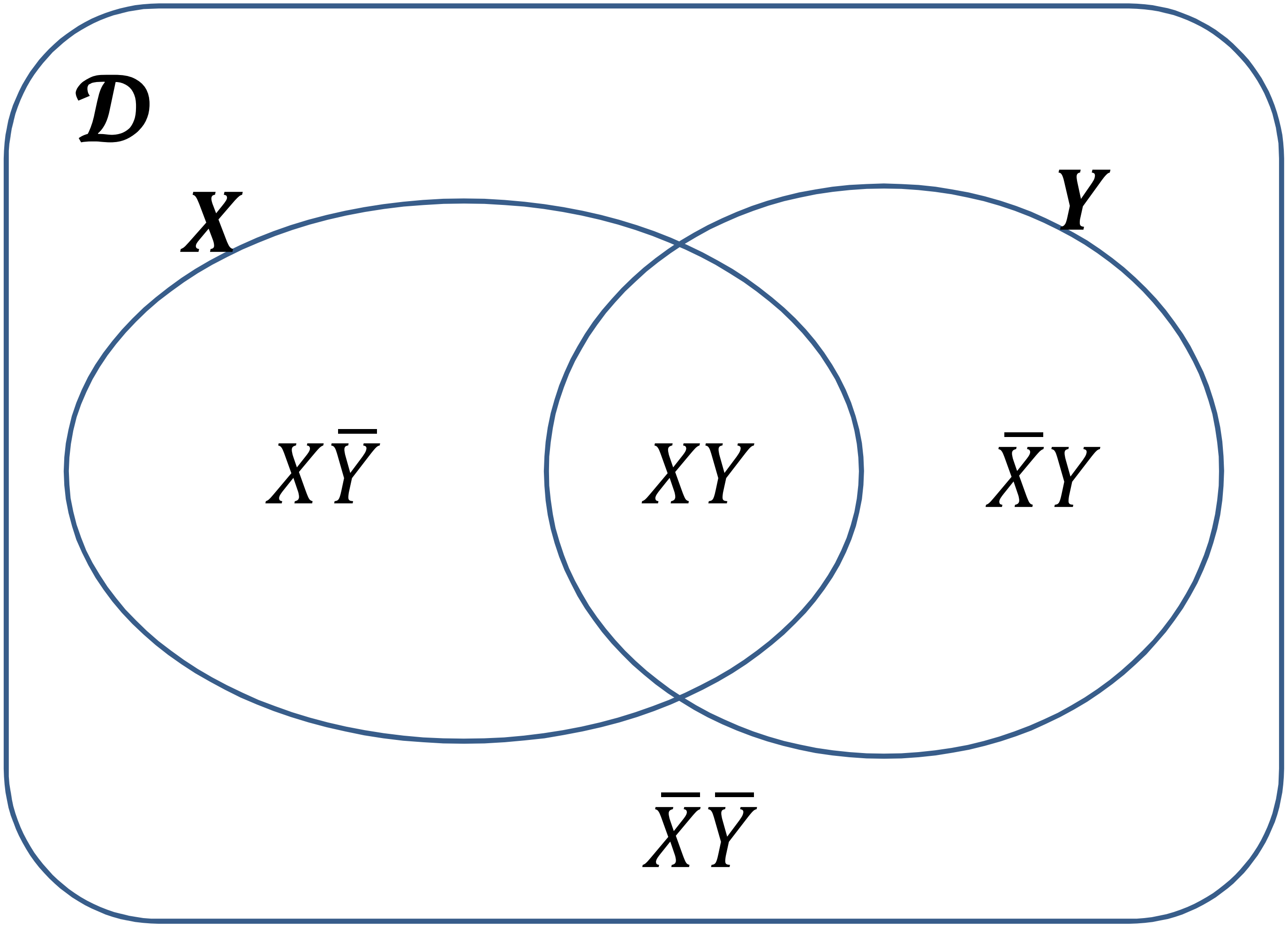
- If we think of the presence/absence (or occurrence/non-occurrence) of Itemsets in a transaction as a Boolean variable, a *Minterm* is a product of these Booleans in which each Boolean appears only once in true or complemented form.
- Example, the Minterms of X and Y are:

$$XY, \bar{X}Y, X\bar{Y}, \bar{X}\bar{Y}.$$

where the bar denotes complement (or negation).

- In general there are 2^n Minterms for n Itemsets.

Minterms for Two Itemsets



Venn Diagram Illustrating the database of transactions $\mathcal{D}$, the Itemsets X and Y, and their Minterms.

Statistical Significance

- The χ^2 Statistic (with one degree of freedom) can be used for testing Dependency at a given *Level of Statistical Significance*, α.
- Usually $\alpha = 1\%$ or $\alpha = 5\%$, the smaller the α the more Statistically Significant the Dependence Rule is.
- Alternatively, we may report the *Confidence Level*, or $1 - \alpha$. For example, for $\alpha = 1\%$ the Confidence Level of the Dependence Rule is 99%.

Minterm Support

- The χ^2 Statistic suffers from a couple of shortcomings:
  - It can be unreliable in the presence of sparsity, which is the case with $\mathcal{D}$;
  - It only tells us statistical significance, not strength of Dependency.
- To mend the first shortcoming, we can use the following robustness criteria for the χ^2 Statistic:
  - All Minterms should have Expected Values of no less than 1;
  - At least 80% of all Minterms should have Expected Values of no less than 5.
- Therefore, we say that an Itemset X has ***Minterm Support*** s at the $p\%$ level if:
  - No Minterm of X has expected value less than 1;
  - At least $p\%$ of the Minterms of X have expected value no less than s.
- Note that if X has Minterm Support $s = 5$ at the $p = 80\%$ level, it would meet the robustness criteria for the χ^2 Statistic.

Strength of Dependency

- The χ^2 Statistic can only tell us a *yes/no answer* of Significance at the $1 - \alpha$ Confidence Level (where $\alpha = p\%$).
- To find a measure of *Strength of Dependency*, we can use information-theoretic arguments. Let's develop those…
- The *Information Gain (or Entropy)* of finding out that an Itemset X is present (occurs) in a transaction is:

$$H(X) = -\sum_{\{minterms\}} P(X) \ln\bigl(P(X)\bigr).$$

- Likewise, the Information Gain from finding that another Itemset Y occurs given that X has occurred (the *Conditional Entropy*) is:

$$H(Y|X) = -\sum_{\{minterms\}} P(Y|X) \ln\bigl(P(Y|X)\bigr).$$

Strength of Dependency

- A measure of the information that X conveys about Y, the *Mutual Information* is:

$$I(X, Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y),$$

- Where $H(X, Y)$ is the *Joint Entropy*:

$$H(X, Y) = -\sum_{\{minterms\}} P(X, Y) \ln\big(P(X, Y)\big).$$

- Note that $I(X, Y) = 0$ if and only if X and Y are independent since in that case $H(Y|X) = H(Y)$:

$$H(Y|X) = -\sum_{\{minterms\}} P(Y) \ln\big(P(Y)\big) \ (Independence).$$

- Moreover, if X and Y are entirely dependent, then $P(Y|X) = 1$, so that $I(X, Y) = H(X) = H(Y)$.

Strength of Dependency

- We are now ready to define the *Dependency Coefficient* as a measure of the Strength of a Dependency Rule of Itemsets X and Y:

$$U(X,Y) = 2\frac{I(X,Y)}{H(X)+H(Y)}.$$

- Note that $U(X,Y)$ is between 0 and 1. It is 0 if X and Y are independent, and it is 1 if they are entirely dependent.
- The extension to k itemsets is:

$$U(X_1,\cdots,X_k) = \frac{k}{k-1}\frac{I(X_1,\cdots,X_k)}{H(X_1)+\cdots+H(X_k)}.$$

- Note that U does not give information about the *directionality* of the dependency. This can be surmised by comparing the measured co-occurrence against the product of the individual occurrences.

Interesting Dependence Rules

- Between the χ^2 Test for Dependence and the Dependency Strength Coefficient we have enough basis to define *Interesting Dependence Rules* to extract useful information from the database of transactions $\mathcal{D}$.
- *An Interesting Dependence Rule $X \Leftrightarrow Y$ occurs when:*
  - *X and Y are dependent and their dependence is statistically significant at the $1-\alpha$ confidence level, and*
  - *The Strength of Dependency is at least some pre-specified number u: $U(X,Y) \geq u$.*

(Aside: What About Power?)

- Null Hypothesis H_0: "*Itemsets are Independent*."

	Null Hypothesis H_0	
	True	False
H_0 Rejected	FP (α)	TP
H_0 Accepted	TN	FN (β)

- Note that we can't measure FP, TP, TN, FN directly because we're dealing with Unsupervised Learning!
- Dependency Framework considers only α.
- Could estimate β (FN) by considering a non-central χ^2 distribution.
- Questions (Project Idea?):
  - Can we come up with a similar rule for β as we did for α?
  - Can we use this rule for pruning?
  - What tradeoffs would there be if the cost of FPs is greater than that of FNs (as is often the case)? (HW)

Tractability

- As in the Support-Confidence Framework, it would be prohibitive to check all possible Itemsets.
- Fortunately, we don't have to check all possibilities.
- To devise a strategy for tractability we have to do some groundwork…
- ***Definition**: Suppose that an Itemset* $X \subset \mathcal{I}$ *possesses a property* $\mathcal{P}$*. If all the subsets of* X *also possess the property* $\mathcal{P}$*, we say that* $\mathcal{P}$ *is <u>Subset Closed</u>,* or *<u>Internally Closed</u> with respect to* X*. If all the supersets of* X *also possess* $\mathcal{P}$*, we say that* $\mathcal{P}$ *is <u>Superset Closed</u>,* or *<u>Externally Closed</u> with respect to* X*.*

Set Closure Complementarity

- **Proposition (Closure Complementarity)**: *An Itemset property $\mathcal{P}$ is Internally Closed if and only if its Boolean complement $\bar{\mathcal{P}}$ is Externally Closed.*
- The proof of this is trivial by noting that if $\mathcal{P}$ is Internally Closed and is true for X, then it must be true, by definition, for all its subsets. On the other hand, if $\mathcal{P}$ is not true for a subset, then it cannot be true for any superset; *i.e.*, $\bar{\mathcal{P}}$ is Externally Closed. A similar argument, along with the identity $\bar{\bar{\mathcal{P}}} = \mathcal{P}$ establishes the remainder of the proof.

Set Closure and Pruning

- Set Closure is useful for pruning Dependency Rules search.
- Suppose a k-Itemset possesses a desirable property (like χ^2 Statistical Dependency) which is Externally Closed. We do not need to check all $(k+1)$-supersets, $(k+2)$-supersets, etc. to see whether they possess this property, since we know they do by virtue of its External Closure.
- Likewise if a k-Itemset possesses an Internally Closed property (like Minterm Support) we do not need to check all its $(k-1)$-subsets, $(k-2)$-subsets, etc. since we know these will also have the property in question.

Internal Closure of Minterm Support

- **Lemma (Minterm Support's Internal Closure)**: *Minterm Support is an Internally Closed Itemset Property.*
- To see this, let X be a k-Itemset with Minterm Support s at the $p\%$ level, and consider, without loss of generality, a $(k-1)$-Itemset $Y \subset X$. We must show that:
  - (1) No Minterm of Y can have expected value less than 1, and
  - (2) at least $p\%$ of Y's minterms have expected value of at least s.
- To show (1), note that we can construct Y by deleting one of the items in X. This means that every minterm in Y can be obtained by the disjoint union of two minterms in X. Therefore, no minterm in Y can have expected value less than 1 by virtue of X's Minterm Support.
- To show (2), note that Y has $2^{k-1}/2^k = \frac{1}{2}$ as many minterms as X. Now, since every minterm in Y is the union of 2 minterms in X, we can see that the fraction of supported minterms in Y can be no less than the fraction of supported minterms in X.

External Closure of χ^2 Statistic

- **Lemma (χ^2 Statistic's External Closure)**: *The χ^2 test for dependence at the $(1-\alpha)$ level of significance is an Externally Closed Itemset Property.*
- The proof is straightforward. It relies on the fact that, regardless of the Itemset size, the χ^2 statistic has 1 degree of freedom, so that all we have to show is that $\chi^2(Y) \geq \chi^2(X)$ for $X \subset Y$.
- The External Closure of the χ^2 statistic tells us that if any Itemset is dependent at the $1-\alpha$ level of significance, then any superset of that Itemset is also dependent at the $1-\alpha$ level of significance.
- Likewise, the Internal Closure that we saw earlier for Minterm Support tells us that if an Itemset has Minterm Support s at the $p\%$ level, then any subset of that Itemset is also Minterm Supported at the same level.

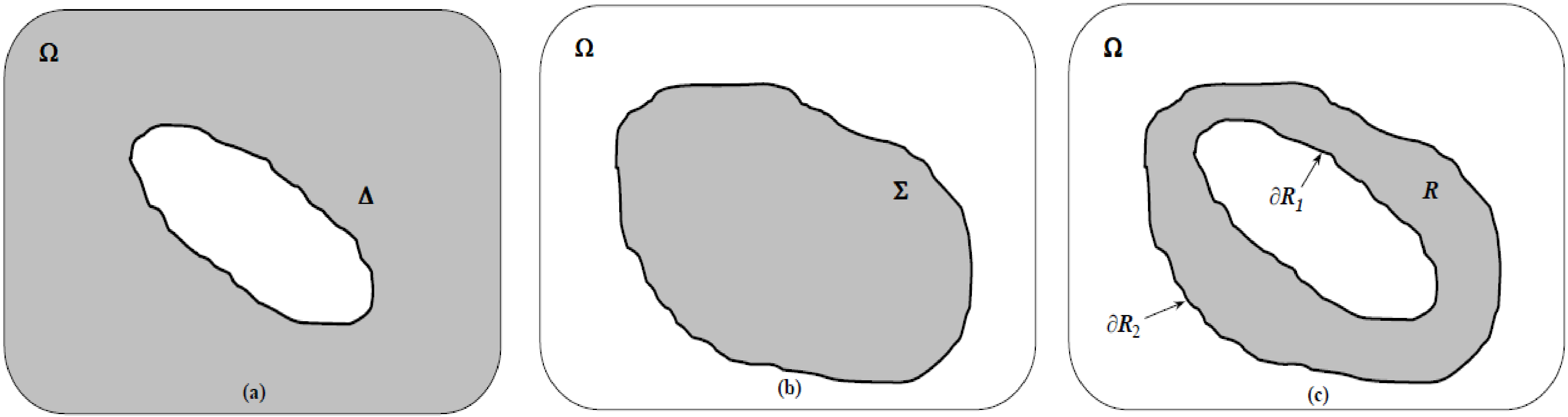
The Region of Interest

- The External Closure of the χ^2 statistic taken together with the Internal Closure of Minterm Support define the *Region of Interest* for the correctly stated Association Rules Problem using the Dependency Framework; *i.e., the reduced search space for Interesting Dependence Rules*.
- Recall that Minterm Support was required to ensure the robustness of the χ^2 test. Therefore, we only consider Itemsets to be interesting if they are Dependent at the $1 - \alpha$ level *and* if they are Minterm Supported with $s = 5$ and $p = 80\%$.
- Closure can be used to prune the search for Interesting Dependence Rules so as to limit ourselves to the much smaller Region of Interest.

The Region of Interest

- To illustrate the Region of Interest, let Σ be the set of all Minterm Supported Itemsets in $\mathcal{J}$, and Δ be the set of all Dependent Itemsets at the $1-\alpha$ Confidence Level, which are supersets of Minterm Supported 2-Itemsets.
- Note that all supersets of Itemsets in Σ are inside Σ, while all subsets of Itemsets in Δ are inside Δ because of the External and Internal Closures of χ^2 and Minterm Support, respectively.
- The Region of Interest is simply the intersection of Σ and Δ.
- If we let Ω represent the set of all possible Itemsets in $\mathcal{J}$ (there are 2^n-1 of these), we can schematically illustrate the relationships between Σ, Δ, Ω, and the Region of Interest…

Schematic of The Region of Interest



Schematic of the Region of Interest where $\mathbf{\Omega}$ represents the Universe of all possible Itemsets.

- (a) The shaded region in (a) represents the External Closure of $\mathbf{\Delta}$, the set of all Dependent Itemsets (of order 2^n).
- (b) illustrates the Internal Closure of $\mathbf{\Sigma}$, the set of all Minterm-Supported Itemsets.
- (c) shows the Region of Interest, $\boldsymbol{R}$ as the intersection of $\mathbf{\Delta}$ and $\mathbf{\Sigma}$. The Internal Boundary $\partial\boldsymbol{R}_1$(the smallest Dependent Itemsets) and the External Boundary $\partial\boldsymbol{R}_2$(the largest Minterm-Supported Itemsets) are also shown. The Region of Interest is bound by $\partial\boldsymbol{R}_1$and $\partial\boldsymbol{R}_2$ and is much smaller than $\mathbf{\Omega}$.

Region of Interest Algorithm

- Our goal is to obtain all Itemsets that:
  - Are Dependent at the $1-\alpha$ level;
  - Are Minterm Supported with $s=5$ and $p=80\%$;
  - Meet the Dependency Strength criterion.
- The first two of these criteria define the Region of Interest.
- We start by determining whether 2-Itemsets are Minterm Supported, and pruning all supersets of the ones that aren't Minterm Supported (using the Internal Closure of Minterm Support along with Closure Complementarity). We can then proceed with 3-Itemsets, etc. until we generate all Minterm-Supported Itemsets.
- Subsequently, we can test for Dependence and prune all subsets of those k-Itemsets that fail the test (using the External Closure of the χ^2 test along with Closure Complementarity).
- The result of this procedure is to generate the *Region of Interest* containing all Interesting Dependence Rules.
- This set of rules can then be ranked by Dependency Strength and those that don't meet a Dependency Strength threshold above a pre-specified value can be pruned.
- (Note that the Dependency Strength Coefficient has *no Closure property* and thus cannot be used for pruning the search except until the very end..)

Itemset Taxonomies

- Itemset “roll-ups” or ***Taxonomies*** can be useful.
- Examples are product categories, classification of stocks by industry, capitalization, etc.
- They can lead to rules such as: “When people buy bakery products they often also tend to buy dairy products.” Or “When Large Cap oil-related equities rise, Midcap manufacturing equities tend to go down.”
- An Itemset *Taxonomy* is defined as a classification scheme where items are grouped into classes. Items are called ***Descendants*** and classes are called ***Ancestors***. A given item can belong to more than one class, but no item can be an Ancestor of any of its ancestors.
- More precisely, we say that an Itemset $\hat{X}$ is an ancestor of Itemset X if all the items in X are either included in $\hat{X}$ or are descendants of items in $\hat{X}$, but X contains no ancestors of any items in $\hat{X}$. (This is known as a *Directed Acyclic Graph*.)
- Note the following two properties relating to Taxonomies:
  - The presence of a descendant in a transaction guarantees the presence of its ancestors.
  - The presence of an ancestor in a transaction does not guarantee the presence of all its descendants.

Taxonomies vs. Sets

- Note the first property is analogous to the relation between a Superset (Descendant) and its Subsets (Ancestors) since the presence of a superset in a transaction guarantees the presence of its subsets.
- However, the second property reverts the analogy since the presence of a Subset does not guarantee the presence of its Supersets.
- This suggests the definition of a "dual" of set closure:
- **Definition**: *Suppose that an Itemset $X \subset \mathcal{I}$ possesses a property $\mathcal{P}$. If all of X's descendants also possess $\mathcal{P}$, we say that $\mathcal{P}$ is <u>Internally Open</u>. If all the ancestors of X also possess $\mathcal{P}$, we say that $\mathcal{P}$ is <u>Externally Open</u>.*

Exploiting Closure-Openness Duality

- Rather than developing new algorithms to exploit Openness for pruning the search space, we can map Openness relations into Closure relations so we can use the algorithms already developed for pruning based on closure.
- **Lemma (Openness Complementarity):** *An Itemset Property $\mathcal{P}$ is Internally Open if and only if $\bar{\mathcal{P}}$ is Externally Open.*
- The proof is easy and follows arguments parallel to those for Closure Complementarity.
- **Definition**: *An Itemset X is said to have a Type I Property, $\mathcal{P}_I$ if that property is non-decreasing in the number of transactions that contain X. On the other hand X is said to have a Type II Property, $\mathcal{P}_{II}$ if that property is non-decreasing in the Cardinality of X (i.e., in the number of classes or groupings in X).*

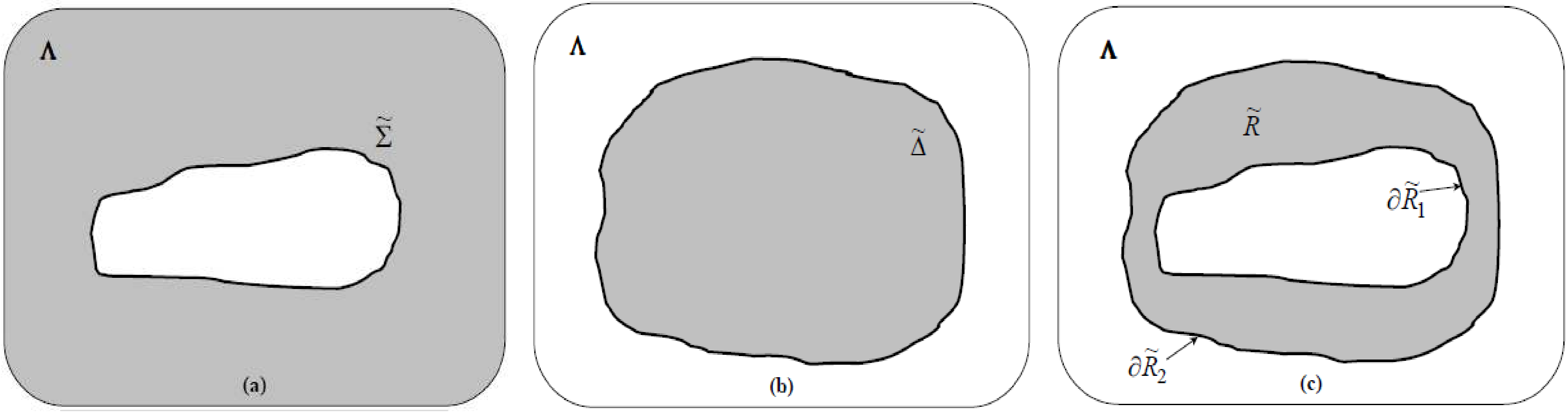
Exploiting Closure-Openness Duality

- An example of a Type I Property is Minterm Support, and an example of a Type II Property is Dependence.
- **Theorem (Closure-Openness Duality)**: *Itemset Properties $\mathcal{P}_I$ and $\mathcal{P}_{II}$ are Externally Open if and only if $\bar{\mathcal{P}}_I$ and $\bar{\mathcal{P}}_{II}$ are Internally Closed. On the other hand, $\bar{\mathcal{P}}_I$ and $\bar{\mathcal{P}}_{II}$ are Internally Open if and only if $\bar{\mathcal{P}}_I$ and $\mathcal{P}_{II}$ are Externally Closed.*
- The proof is straightforward and left as an exercise. The following Corollaries follow immediately:
  - *Let $\hat{X}$ be an ancestor of X according to some Taxonomy. If $\hat{X}$ is an independent Itemset, then X is also independent. Conversely, if X is dependent then $\hat{X}$ is also dependent.*
  - *If X is Minterm Supported then so is $\hat{X}$.*
  - *If X is χ^2-Dependent at the $1 - \alpha$ level, then so is $\hat{X}$.*

Exploiting Closure-Openness Duality

- It then follows that we can apply the Closure-Pruning algorithms that we saw before to Taxonomies by making the following identifications:
  - We treat Ancestors as Subsets when it comes to Minterm Support (which is Externally Open and Internally Closed);
  - We treat Ancestors as Supersets when it comes to Dependence (which is Externally Open and Externally Closed).
- The following Figure illustrates the Openness Relations and the Taxonomy Region of Interest…

Taxonomy Region of Interest Schematic



Schematic of the Taxonomic Region of Interest where $\boldsymbol{\Lambda}$ represents the Universe of all possible Itemsets with respect to a Taxonomy.

- (a) The shaded region in (a) represents the External Openness of $\widetilde{\boldsymbol{\Sigma}}$, the set of all Dependent Itemsets.
- (b) illustrates the Internal Openness of $\widetilde{\boldsymbol{\Delta}}$, the set of all Minterm-Supported Itemsets.
- (c) shows the Region of Interest, $\widetilde{\boldsymbol{R}}$ as the intersection of $\widetilde{\boldsymbol{\Delta}}$ and $\widetilde{\boldsymbol{\Sigma}}$. The Internal Boundary $\partial\widetilde{\boldsymbol{R}}_1$ and the External Boundary $\partial\widetilde{\boldsymbol{R}}_2$ are also shown. The Region of Interest is bound by $\partial\widetilde{\boldsymbol{R}}_1$ and $\partial\widetilde{\boldsymbol{R}}_2$ and is much smaller than $\boldsymbol{\Lambda}$.

Efficient Algorithms: Sampling

- We've considered algorithms that emphasize both correctness and computational tractability. We now focus on improving computational efficiency while paying a small price in correctness.
- We would like to manipulate the database of transactions in resident memory. To do so, we use a *Sample Database* $\mathcal{S}$ instead of the full database $\mathcal{D}$, and consider the *impact on correctness*.
- We take a random sample $\mathcal{S}$ of size m_s *with replacement* from the original database $\mathcal{D}$, and consider the error on Minterm Support introduced by the sampling.

Efficient Algorithms: Sampling

- Suppose we have an Itemset X with “true” Minterm Support $s(X)$ (in the full database $\mathcal{D}$), while the *sample* Minterm Support is $s(X, m_s)$ (in the sample database $\mathcal{S}$). We define the ***Sampling Error*** as:

$$\mathcal{E}(X, m_s) = |s(X) - s(X, m_s)|.$$

- If m_s is chosen such that

$$m_s \geq \frac{1}{2\epsilon^2} \ln\left(\frac{2}{\delta}\right), then$$

$$P(\mathcal{E}(X, m_s) > \epsilon) \leq \delta.$$

- In other words, the probability that the sampling error is greater than some number ϵ is at most δ.
- This follows from the fact that we can think of the sampled Itemset as a binomially distributed random variable with “true” probability of success $s(X)$, and applying the ***Chernoff Bound***.

Efficient Algorithms: Sampling

- The previous bound relates the sample size to the support error. We must trade off the error against being able to fit the Sample Database in main memory.
- Suppose we choose a sample size given the memory constraints. We can reduce the probability of error given the sample size if we relax the minimum acceptable support. The tradeoff here is that we will generate more Itemsets that would have to be tested for meeting the desired Support threshold.
- If we have a Required Support s and we wish to specify the bound on the probability of sampling error to some δ, we must reduce the support threshold to a new support threshold s_δ bounded by:

$$s_\delta < s - \sqrt{\frac{1}{2m_s} ln\left(\frac{1}{\delta}\right)}.$$

- This follows from applying the *one-sided Chernoff Bound*.

Efficient Algorithms: Sampling

- The Sampling Algorithms work as follows. We use a sample database $\mathcal{S}$ of size m_s obtained by sampling the full transaction database $\mathcal{D}$ randomly and with replacement.
- We run the Region of Interest algorithms on the (smaller) sample database $\mathcal{S}$. This will produce some Rules (Itemsets) that are not interesting. These Itemsets can be deleted by checking them against the full database $\mathcal{D}$ at the end of the procedures. However, there is a chance that we could miss some Itemsets that would be interesting had we run the algos on $\mathcal{D}$ to begin with.
- In picking m_s such that it is small enough to fit in main memory we have two choices. First, we can keep the required Support and accept the error. Second we can lower the required Support threshold and reduce the error bound, but we would increase the search time by considering Itemsets that may turn out to be unsupported.

Efficient Algorithms: Partitioning

- Instead of sampling the database, we could Partition it into subsets that are small enough that they fit in main memory, process each partition, and then combine the end results (this makes it *parallelizable*).
- In other words, we form a *Partition* $\mathcal{P}_N$ of $\mathcal{D}$ by taking N samples of size m_p from $\mathcal{D}$, such that $Nm_p = m$. The samples are chosen randomly but without replacement so that they are disjoint but span $\mathcal{D}$.
- For each of the partitions we generate *local candidate Itemsets* that satisfy local Minterm Support.
- The key here is that the set of all globally Minterm-Supported Itemsets will be a subset of the union of all locally Minterm-Supported Itemsets (why?).
- All the candidate Itemsets generated using local Support in each of the partitions will contain all of the “true” Supported Itemsets in the full database $\mathcal{D}$.

Efficient Algorithms: Partitioning

- More formally, let Σ be the set of all globally Minterm-Supported Itemsets obtained from the full database $\mathcal{D}$, and let Σ_i be the set of locally Minterm-Supported Itemsets in the i^{th} element of a partition $\mathcal{P}_N$ of $\mathcal{D}$. For any Itemset X, we have:

$$X \subset \Sigma \Rightarrow X \subset \bigcup_{i=1}^{N} \Sigma_i.$$

- In other words, if X is globally Minterm-Supported, then X must be contained in at least one of the locally Minterm-Supported elements of the partition $\mathcal{P}_N$ of $\mathcal{D}$. (The proof is beyond our scope.)
- The procedure for the partitioning scheme is to generate a superset of all globally Supported Itemsets using partition elements that are small enough to fit in main memory. Then, this superset is checked against the full database $\mathcal{D}$, and the un-supported Itemsets are discarded.

Association Rules Conclusion

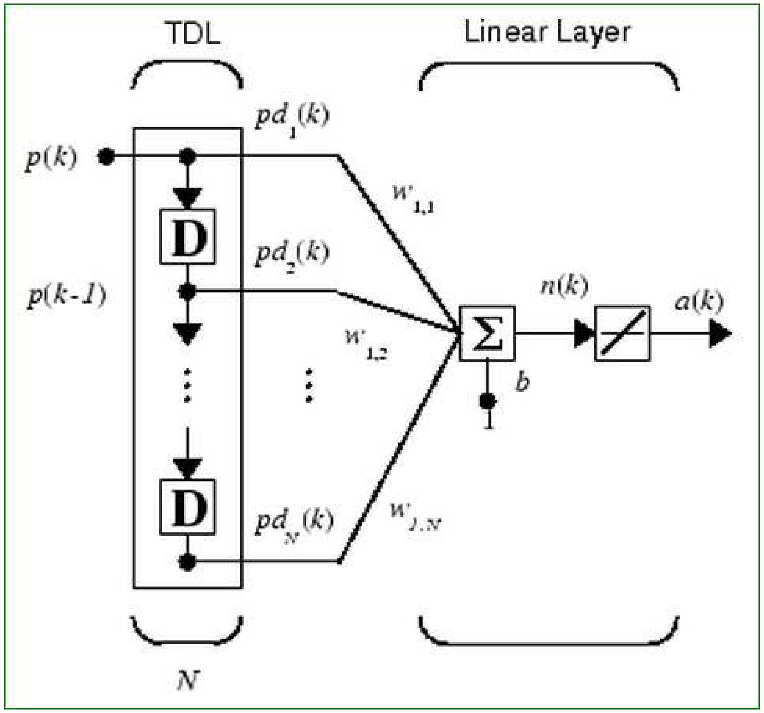
- The Association Rule Problem using the Support-Confidence Framework *ignores statistical significance* and can produce spurious rules and discard significant ones.
- The Dependency Framework does take into account statistical significance and *distinguishes useful rules from random ones*.
- The search space for Association Rules is prohibitively large, so we must exploit Closure Relations to limit the search to a much smaller Region of Interest in order to make the problem tractable.
- We can also use Sampling and Partitioning of the Transaction Database to make the problem even more computationally expedient.
- For Taxonomies we can exploit Openness Relations that can be mapped to Closure Relations through a duality property. This in turn allows pruning of taxonomies through Closure Relations.

Dynamic ANNs and Time Series Prediction

- A time series is a sequence of stochastic variables ordered in time:

  $$y_1, y_2, y_3, \dots$$

  where the subscript is a time index.
- We can use ANNs to approximate a time series by treating the time series as a single input pattern, but introducing Delays ($\mathbf{D}$) via a *Tapped Delay Line* (TDL):

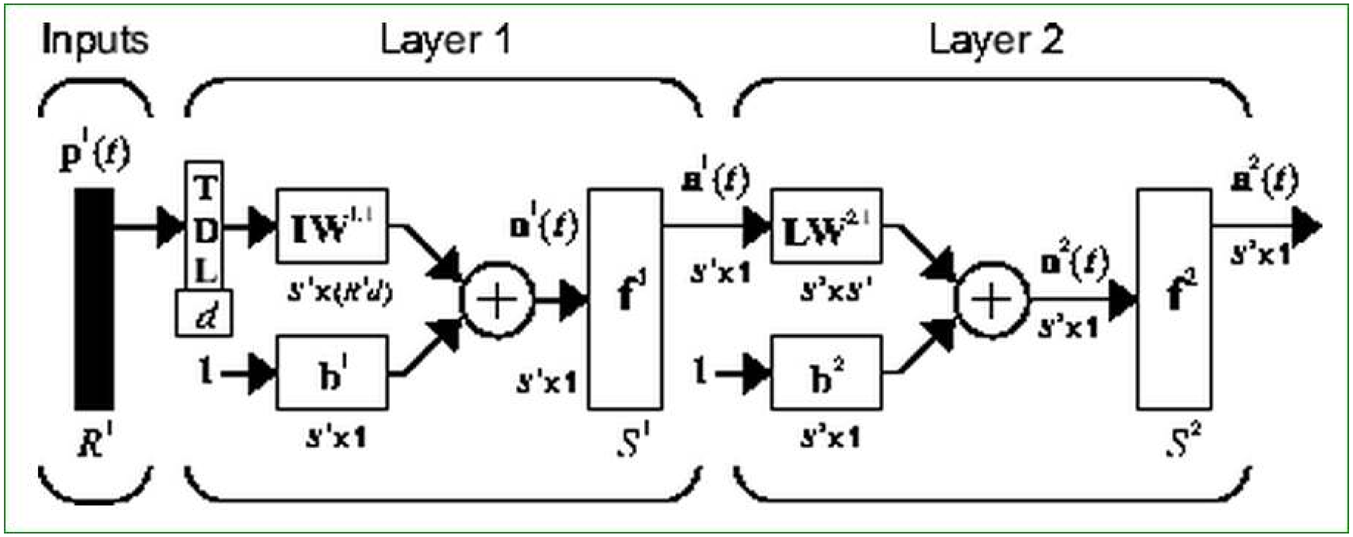


- This shows a TDL with N delays feeding into a single ADALINE. This can be used for approximating linear autoregressive (AR) time series:

$$y_t = \sum_{k=1}^{N} \varphi_k y_{t-k} \,.$$

Dynamic ANNs and Time Series Prediction

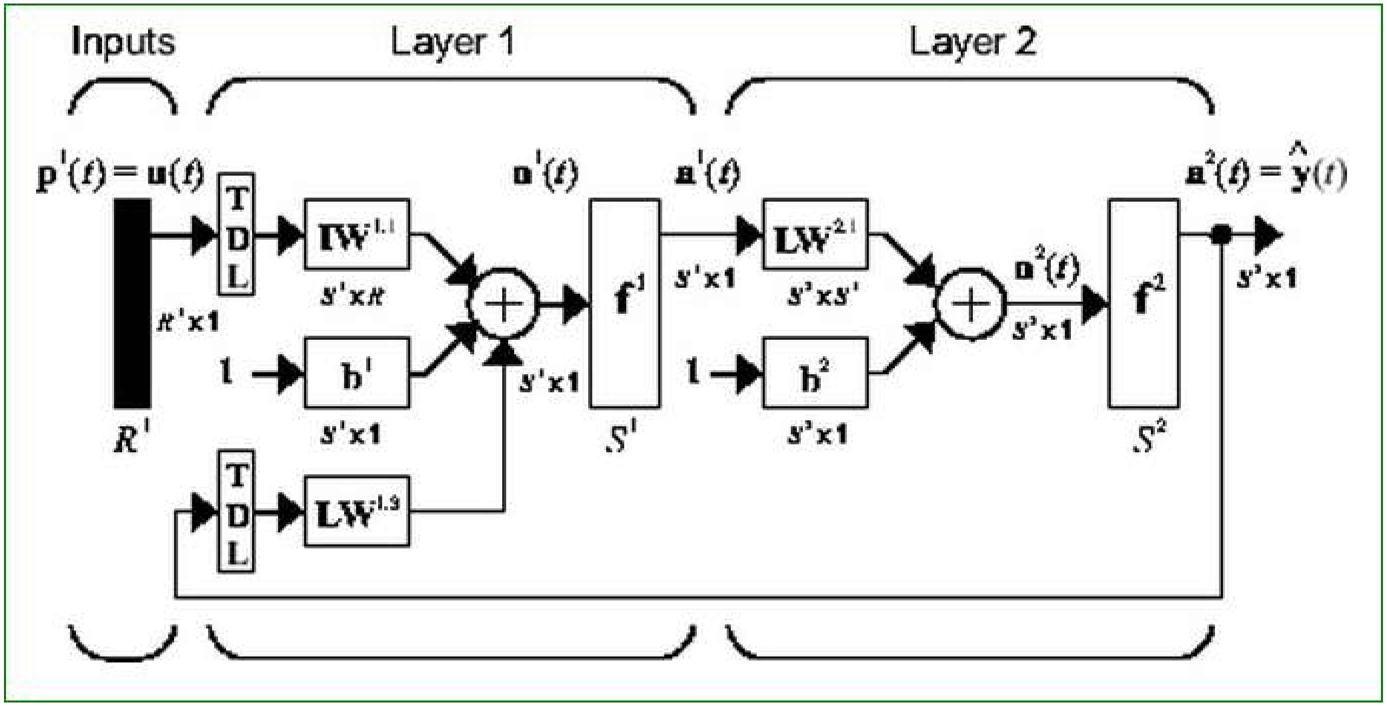
- To generalize this linear filter, we can add *multiple, non-linear hidden nodes* for prediction of a more general time series:



- Note that: Inputs $\mathbf{p}^1$ are a (multi-dimensional) time series;
- Inputs are presented sequentially but delayed *d* time steps by the TDL;
- The Activation Functions $\mathbf{f}^1$ and $\mathbf{f}^2$ can be nonlinear, particularly the Hidden-Layer Activation $\mathbf{f}^1$. (Usually is $\mathbf{f}^1$ sigmoidal and is $\mathbf{f}^2$ is linear.);
- The network does not start producing outputs until the TDL is full, *i.e.*, until *d* inputs have been presented. Thereafter, the network produces sequential outputs every time an input is presented;
- This is a *Feedforward Architecture with a TDL* where the dynamics (delays) happen only in the input layer. This is also known as an *Open Architecture*.

Dynamic ANNs and Time Series Prediction

- In ***Recurrent Networks***, the output is fed back into the input via a TDL:



- This is also known as a ***Closed Architecture*** or ***Closed-Loop Architecture***.
- Note that the inputs can include the output series $\boldsymbol{y_t}$ itself, and/or a "*Companion*" or "*Regressor*" series $\boldsymbol{x_t}$.
- This can approximate a more general (possibly non-linear) series:

$$\boldsymbol{y_t} = f(\boldsymbol{y_{t-1}}, \boldsymbol{y_{t-2}}, \dots, \boldsymbol{y_{t-m}}, \boldsymbol{x_{t-1}}, \boldsymbol{x_{t-2}}, \dots, \boldsymbol{x_{t-n}}),$$

  where f is the (possibly non-linear) function to be approximated.
- For example, the target can be "Exchange Rate" and regressors can be "Oil Price," "T-Bill Rates" etc.

Questions/Comments

- Project questions/ideas?