Dataset Shifts in Autonomous Systems

CIS700: Safe Autonomy

James Weimer

February 5, 2019

Penn Engineering

PRECISE

PENN RESEARCH IN EMBEDDED COMPUTING AND INTEGRATED SYSTEMS ENGINEERING

Outline

- Reading Material Recap
  - 2-3 minute impromptu overview
- What are Dataset Shifts?
  - Examples in medicine and anomaly detection
- Types of Dataset Shifts
  - Covariate shifts, prior probability shifts, concept drift
- Common causes of Dataset Shifts
  - Can these be used to improve detection?
- Common Assumptions in Dataset Shift Detection
  - What needs to be true to perform dataset shift detection?
- Research challenges in Dataset Shifts

Student Recap of Reading Material

- Sugiyama, Masashi, Neil D. Lawrence, and Anton Schwaighofer. *Dataset shift in machine learning*. The MIT Press, 2017.
  - Student/co-instructor: Jim Weimer
- Moreno-Torres, Jose G., et al. "A unifying view on dataset shift in classification." *Pattern Recognition* 45.1 (2012): 521-530.
  - Student: Ramneet Kaur
- Raza, Haider, Girijesh Prasad, and Yuhua Li. "EWMA based two-stage dataset shift-detection in non-stationary environments." *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, Berlin, Heidelberg, 2013.
  - Student: Tyler Oliveri
- Klinkenberg, Ralf, and Thorsten Joachims. "Detecting Concept Drift with Support Vector Machines." *ICML*. 2000.
  - Student: Kyle Leonard

Class Poll

- We monitor shifts in datasets as a proxy for classifier performance?

- Which is more susceptible to test data not matching training data, SVMs or DNNs?
  – e.g., SVMs are simple linear classifiers, DNNs are more complex

- Is end-to-end control a good or bad idea? Why?
  – e.g., avoid feature engineering – just train a model to produce actuation commands.

What are Dataset Shifts?

- “… data experience a phenomenon that leads to a change in the distribution …”
- “… the joint distribution of inputs and outputs differs between training and test stage.”

- Space suffers from lack of standardized terminology
  - concept shift, changes of classification, changing environments, contrast mining, fracture points, fractures between data, sample bias, …
  - Very hard to find solutions to existing problems

Notation and Terminology

- x are covariates
  - e.g., inputs, features, raw data, independent variables (?)…
- y are targets
  - e.g., outputs, labels, classes, dependent variables (?) …

- $p(x, y)$ is the joint distribution when training
- $p'(x, y)$ is the joint distribution when testing

- Bayes' Law
  - $p(x, y) = p(y|x)p(x) = p(x|y)p(y)$
  - $p'(x, y) = p'(y|x)p'(x) = p'(x|y)p'(y)$

Causation vs. Correlation

- Lots of confusion about correlation and causation in data …
  - Causality is an intrinsic property of the data generation process
  - Correlation is a relative property of the data
- **Jim's Opinion:** *Safety critical autonomy should only consider causal features and labels*
  - *Just because two things happen, doesn't mean they are (or are not) related*
  - Causation is usually a direct consequence of the covariate and label choices
- Two causal scenarios:
  - $X \rightarrow Y$ problems: labels are causally determined by covariates
    - **Anomaly Detection:** Detecting anomalous accelerometer data in cars
  - $Y \rightarrow X$ problems: labels causally determine covariates
    - **Smart Alarms in Medicine:** Silencing false heart rate alarms

This is why I don't like calling X independent and Y dependent variables!!!

Example 1: Medical Smart Alarms

- Patient movement is a problem for many medical devices
  - Leads to false alarms
- Lets build a detector to determine whether a patient is moving
- Defining features and labels
  - Heart Rate can be measured by a Pulse Oximeter and ECG independently
    - Lets define a feature to be the difference in heart rate
      - $x = |HR_{PO} - HR_{ECG}|$
  - Often when the patient moves, pulse ox is wrong
    - Lets define a label to be whether the patient is moving
      - $y = 0$: patient is still
      - $y = 1$: patient is moving
- Are they related causally? How?
  - Movement influences pulse oximeter heart rate, $Y \rightarrow X$
    - *"labels causally determine covariates"*
- What if we wanted to monitor a doctors response to alarms?

You must understand your data generation process!!!

Example 2: Anomaly Detection

- Accelerometer data can be affected by all sorts of disturbances
  - May lead to bad decisions later if anomalies are not detected.
- Lets build a detector to determine whether the accelerometer data is anomalous
- Defining features and labels
  - Features are the recent accelerometer data measurements
    - x is the average of the last 10 measurements
  - Labels are whether the data is flagged to be anomalous
    - $y = 0$: not anomalous
    - $y = 1$: anomalous
- Are the features and labels related causally? How?
  - The features, representing recent behavior determines whether data is anomalous -- i.e., $X \rightarrow Y$
  - *``labels are causally determined by covariates"*
- What if we wanted to detect a bias in the accelerometer?

You must understand your data generation process!!!

Types of Dataset Shifts

- Dataset shift defined as $p(x, y) \neq p'(x, y)$
  - In general too hard to do anything … good luck
- Subclasses of dataset shift
  - Covariate shift:
    - $p(y|x) = p'(y|x)$ and $p(x) \neq p'(x)$ in $X \rightarrow Y$ problems
  - Prior probability shift:
    - $p(x|y) = p'(x|y)$ and $p(y) \neq p'(y)$ in $Y \rightarrow X$ problems
  - Concept shift
    - $p(y|x) \neq p'(y|x)$ and $p(x) = p'(x)$ in $X \rightarrow Y$ problems
    - $p(x|y) \neq p'(x|y)$ and $p(y) = p'(y)$ in $\mathrm{Y} \rightarrow X$ problems
- Common causes of dataset shift
  - **data generation:** sample selection bias, missing data, etc.
  - **non-stationary environments**: seasonal changes, location, etc.
- Lots of literature on cause-specific dataset shift
  - more information = better detection

Examples to follow on all these

Could be a nice class project …

Outline

- Reading Material Recap
  - 2-3 minute impromptu overview
- What are Dataset Shifts?
  - Examples in medicine and anomaly detection
- Types of Dataset Shifts
  - Covariate shifts, prior probability shifts, concept drift
- Common causes of Dataset Shifts
  - Can these be used to improve detection?
- Common Assumptions in Dataset Shift Detection
  - What needs to be true to perform dataset shift detection?
- Research challenges in Dataset Shifts

Examples Revisited – Building Classifiers

- Medical Smart Alarms
  - Detect movement from HR variability in PulseOx and ECG
    - $x = |HR_{PO} - HR_{ECG}|$
    - $y \in \{0,1\}$
      - 0 = no movement
      - 1 = movement
  - Build a classifier
    - Collect training data: (x, y)
      $\{(.5,0), (.4,0), (.6,0), (1.9,1), (2.0,1), (2.1,1)\}$
    - Train a classifier
      - $\hat{y}(x)$: $x \leq 1 \leftrightarrow y = 0$
    - Perform testing and validation
  - What are the potential dataset shifts?

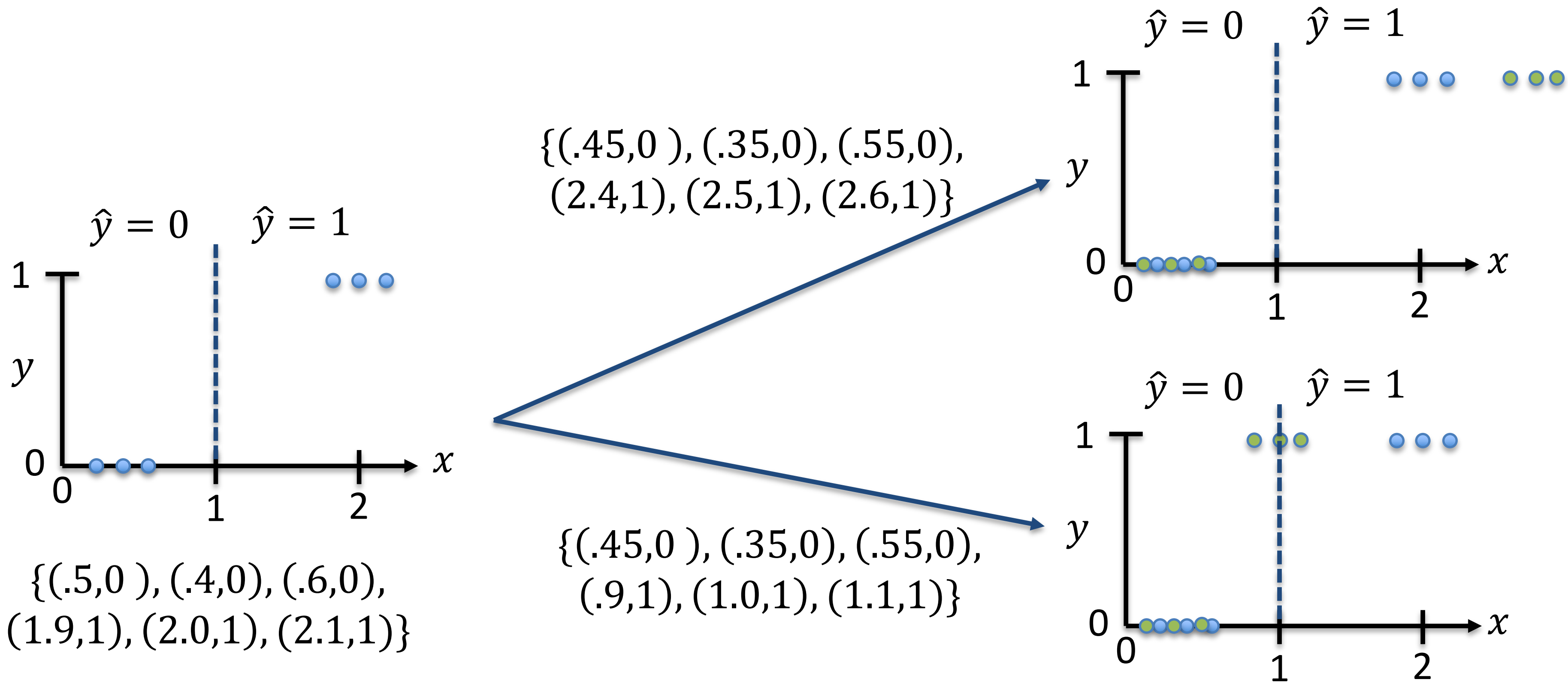
$Y \rightarrow X$ problem:
Prior distribution shift
Concept shift

- Anomaly Detection
  - Detect anomaly from accelerometer data
    - $x = \frac{1}{N} \sum_{n=1}^{N} |acc_n - m|$
    - $y \in \{0,1\}$
      - 0 = no anomaly
      - 1 = anomaly
  - Build a classifier
    - Collect training data
      - $\{(.5,0), (.4,0), (.6,0), (1.9,1), (2.0,1), (2.1,1)\}$
    - Train a classifier
      - $\hat{y}(x)$: $x \leq 1 \leftrightarrow y = 0$
    - Perform testing and validation
  - What are the potential dataset shifts?

$X \rightarrow Y$ problem:
Covariate shift
Concept shift

Covariate Shift

- Recall our definition for covariate shift:
  - $p(y|x) = p'(y|x)$ and $p(x) \neq p'(x)$ in $X \to Y$ problems



What properties make a good covariate shift detector?

Detecting Covariate Shifts

- What makes a good covariate shift detector?
  - Relies predominantly on features (not labels)
  - Minimal additional assumptions

- Common tests used for Covariate Shifts
  - Parametric tests:
    - Typically use assumptions on distribution of sample average/variance
      - Student's t-test, F-test, Chi-squared test, etc.
    - If assumptions hold, can be optimal
  - Nonparameteric tests:
    - Minimal assumptions on underlying distributions
      - Wilcoxon Rank-Sum, Kolmogorov-Smirnov test, etc.

- Best tests tend to use techniques from both
  - Combine a parametric test as a trigger for a nonparametric test.
    - Raza, Haider, Girijesh Prasad, and Yuhua Li. "EWMA based two-stage dataset shift-detection in non-stationary environments." *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, Berlin, Heidelberg, 2013.

Some Parametric Tests for Covariate Shifts

- Student's T-test: Given population mean μ
  - $t = \frac{\bar{x}-\mu}{s\sqrt{N}}$, $\bar{x} = \frac{1}{N}\sum x_n$, $s = \sqrt{\frac{1}{N}\sum (x_n - \bar{x})^2}$,
  - Pros: asymptotically optimal when sample mean, $\bar{x}$, equals μ
    - Provides tight confidence estimate/interval
  - Cons: susceptible to selection bias in data, requires known μ
  - Lots of variations on this test – good starting point.

- F-test / Analysis of Variance – ANOVA:
  - F = explained variance / unexplained variance
    - One examples: $F = t^2$
  - Pros: good for comparing statistical models fitted to data
    - i.e., learned models
  - Cons: more data needed than a T-test, high variance results in low testing power.
  - ANOVA is widely used in situations where testing conditions can be controlled

Some Nonparametric Tests for Covariate Shifts

- Wilcoxon Rank Sum Test:
  - Tests whether two samples are from populations with the same distribution
    - Given two samples (test and training data), test whether a random sample from one is greater than the other
  - Pros: non-parametric version of T-test, almost as powerful
  - Cons: requires features, x, to be ordinal
    - Can introduce bias in multiple dimensions
  - Variants include: Wilcoxon signed-rank test
- Kolmogorov-Smirnov Test:
  - Tests whether data came from the same distribution
    - Tests whether the cdf's are the same
  - Pros: Sensitive to location and shape, very general and useful
  - Cons: requires features, x, to have a single dimension
  - See also Anderson-Darling and Shapiro-Wilk test

Hybrid Tests for Covariate Shifts

- Why a hybrid test?
  - Parametric tests use data more efficiently
  - Nonparametric tests are robust to assumptions of parametric tests
- A common approach:
  - Use a general model to filter data to produce i.i.d. measurements (null hypothesis)
    - Apply chi-square, t-test, or F-test, etc.
  - If null hypothesis is rejected, trigger nonparametric test
    - Apply Wilcoxon Rank Sum, Kolmogorov Smirnov, etc.

EWMA Covariate Shift Detector

Input: Submit the training dataset to the training phase and compute the parameters for testing.
Receive new data in the testing phase sample-by-sample and perform the check as follows.
IF (Shift detected)
THEN (Report the point of shift and move to stage-II for validation)
ELSE (Continue and integrate the upcoming information).
Output: Shift-detection points.

Stage-I
Training Phase
1. Assign training data to $x_{(i)}$ for $i=1:n$, where n is the number of observations in training data
2. Calculate the mean of $x_{(i)}$ and set as $z_{(0)}$.
3. Compute the z-statistics for each observation $x_{(i)}$ in training data for a range of λ values.
   $z_{(i)} = \lambda x_{(i)} + (1-\lambda) z_{(i-1)}$
4. Estimate λ by minimizing over the training dataset the square of 1-step-ahead prediction error: $err_{(i)} = x_{(i)} - z_{(i-1)}$.
5. Finally estimate the variance of error for the testing phase.

Testing Phase
1. For each data point $x_{(i)}$ in the operation/testing phase
2. Compute $z_{(i)} = \lambda x_{(i)} + (1-\lambda) z_{(i-1)}$
3. Compute $err_{(i)} = x_{(i)} - z_{(i-1)}$
4. Estimate the variance $\hat{\sigma}^2_{err_{(i)}} = \vartheta\, err^2_{(i)} + (1-\vartheta)\hat{\sigma}^2_{err_{(i-1)}}$
5. Compute $UCL_{(i)}$ and $LCL_{(i)}$:
6. $UCL_{(i)} = z_{(i-1)} + L\hat{\sigma}_{err_{(i-1)}}$
7. $LCL_{(i)} = z_{(i-1)} - L\hat{\sigma}_{err_{(i-1)}}$
8. IF ($LCL_{(i)} < x_{(i)} < UCL_{(i)}$)
   THEN (Continue processing)
   ELSE (Go to Stage-II)

Stage-II
1. For each $x_{(i)}$
2. Wait for m observations after the time i, organize the sequential observations around time i into two partitions, one containing $x_{((i-(m-1)):i)}$, another $x_{((i+1):(i+m))}$.
3. Execute the hypothesis test on the partitioned data
4. IF ($H=1$)
   THEN (test rejects the null hypothesis): Alarm is raised
   ELSE (The detection received by stage-I is a false and discarded)

Design Time:
1) Assign order to training data
2) Fit model to data
   - i.e., pick model parameters
3) Calculate error and variance
   - i.e. pick test threshold

Run Time (Phase 1):
1) Assign order to test data
2) Apply model to test
3) Compare with threshold
   1) It alarm, GOTO Phase 2

Run Time (Phase 2):
1) Divide data into pre-alarm and post-alarm.
   - based on Phase 1
2) Check using non-parametric test

Covariate Shift Detection Summary

- Covariate shifts are on the features
  - Features are observable online
- Not all covariate shifts cause the classifier to fail
  - In fact, classifier performance has nothing to do with detecting covariant shifts
- Key insights to designing monitors for covariate shifts
  - Understand your data generation process
    - Do your labels depend on your features? (The answer has to be yes)
      - If no, then you may be able to change your problem …
  - Make sure your assumptions are valid
    - Parametric vs. non-parametric tests

Potential Covariate Shift Research Directions

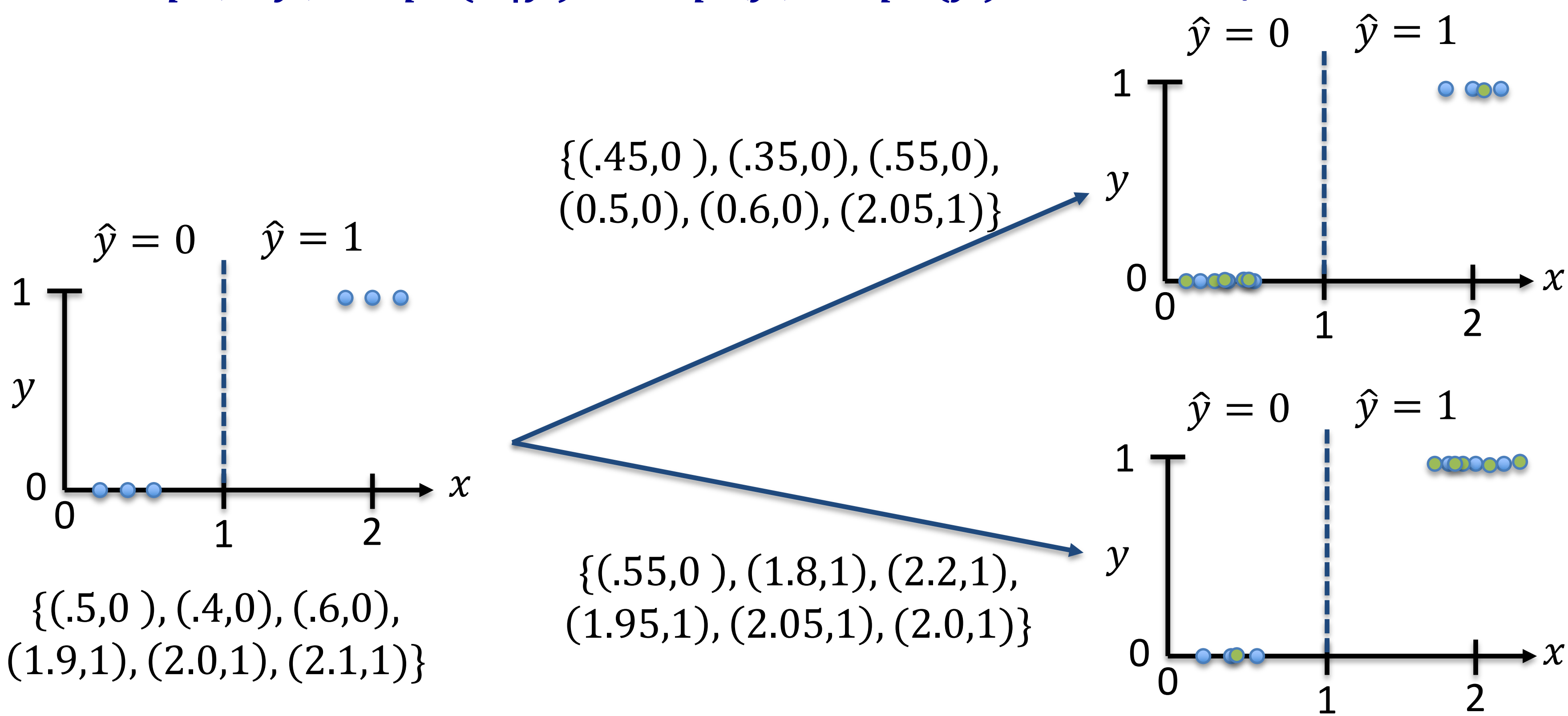
- Apply Covariate shift detectors to existing classifiers
  - Talk to us for a list of potential classifiers
- Utilize more complex models of data generation dynamics
  - Autoregressive moving average with exogeneous inputs (ARMAX)
  - Autoregressive integrated moving average (ARIMA)
- Identify nuisances in underlying data generation process and remove their effect from features post-training
  - Especially interesting if the trained classifier is also invariant to nuisances
- Online quantification of covariate shift detection performance
  - Add nuisances to the features (induce covariate shifts) and monitor the ability to detect them.

Outline

- Reading Material Recap
  - 2-3 minute impromptu overview
- What are Dataset Shifts?
  - Examples in medicine and anomaly detection
- Types of Dataset Shifts
  - Covariate shifts, **prior probability shifts**, concept drift
- Common causes of Dataset Shifts
  - Can these be used to improve detection?
- Common Assumptions in Dataset Shift Detection
  - What needs to be true to perform dataset shift detection?
- Research challenges in Dataset Shifts

Prior Probability Shift

- Recall our definition for prior probability shift:
  - $p(x|y) = p'(x|y)$ and $p(y) \neq p'(y)$ in $\mathrm{Y} \rightarrow X$ problems



When is a shift in prior probability a concern?

Detecting Prior Probability Shifts

- When are prior probability shifts a concern?
  - When priors on the labels affect classifier performance
  - When does this happen? Baysian techniques for learning.
    - More precisely, anytime the frequency of labels affects what is learned.

- How to detect prior probability shifts?
  - Scenario 1: We can request labels online
    - e.g., ask for intermittent labeling from an Oracle
  - Scenario 2: We can not request labels online
    - Only features are observed – not labels
  - to optimal testing for distributions with discrete support

- Scenario 1: Requesting labels at runtime
  - Active area of research:
    - semi-supervised learning, active learning, etc.
  - Most techniques are devoted to minimizing the need for labels to perform learning
    - Keep in mind: you can always do better with more information
    - Only makes sense in autonomy if labeling incurs high cost
  - For prior probability shifts this allows us to pursue different classification techniques
    - Many labels are discrete, allows for distribution monitoring
      - Another active area of research devoted to optimal testing for distributions with discrete support

Detecting Prior Probability Shifts

- Scenario 2: We can not observe the labels online
  - All the techniques used for covariate shifts can be applied
    - Need to assume: $p(x) \neq p'(x) \leftrightarrow p(y) \neq p'(y)$
      - assumes that the prior probability shift affects feature distribution
      - If features are informative, this is a property of the data generation process

- What if the prior probability shift doesn't affect the feature distribution?
  - You have performed a poor feature selection
  - Performance could be significantly different than expected

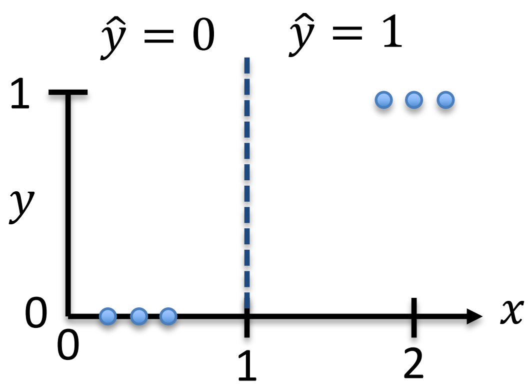
- If we are just going to use the same techniques – why distinguish between covariate shifts and prior probability shifts?
  - Because it requires an additional assumption
  - **Requires a more restrictive data generation process**

Outline

- Reading Material Recap
  - 2-3 minute impromptu overview
- What are Dataset Shifts?
  - Examples in medicine and anomaly detection
- Types of Dataset Shifts
  - Covariate shifts, prior probability shifts, **concept drift**
- Common causes of Dataset Shifts
  - Can these be used to improve detection?
- Common Assumptions in Dataset Shift Detection
  - What needs to be true to perform dataset shift detection?
- Research challenges in Dataset Shifts

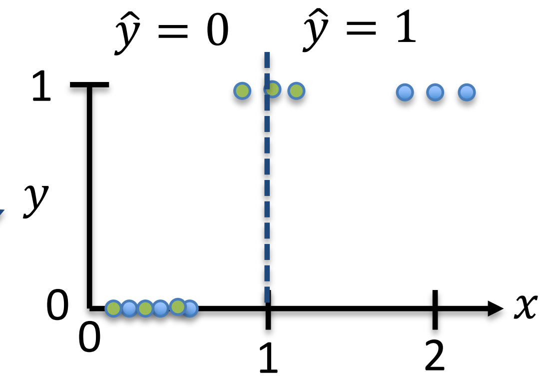
Concept Shift

- Recall our definition for concept shift:
  - $p(x|y) \neq p'(x|y)$ and $p(y) = p'(y)$ in $\mathrm{Y} \rightarrow X$ problems
  - $p(y|x) \neq p'(y|x)$ and $p(x) = p'(x)$ in $\mathrm{X} \rightarrow Y$ problems

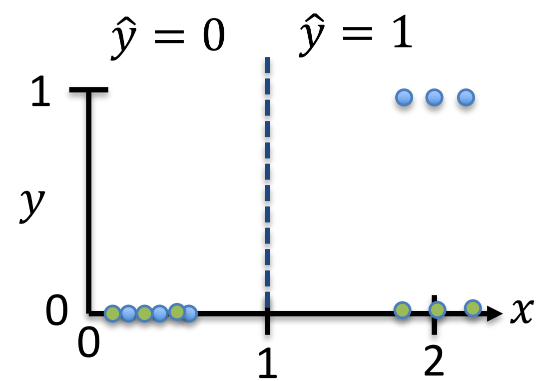


$\{(.5,0), (.4,0), (.6,0), (1.9,1), (2.0,1), (2.1,1)\}$

$\mathrm{Y} \rightarrow X$ problem
$\{(.45,0), (.35,0), (.55,0), (.9,1), (1.0,1), (1.1,1)\}$



$\mathrm{X} \rightarrow Y$ problem
$\{(.45,0), (.35,0), (.55,0), (2.05,0), (1.95,0), (2.0,0)\}$



How can we detect concept drift?

Detecting Concept Drift

- Detecting concept drift is the most challenging
- In $X \rightarrow Y$ problems: concept drift is unobservable
  - Assumes same feature distributions
    - Can not be detected without test data labels – Can we use a labeler?
- In $Y \rightarrow X$ problems: concept drift has low specificity
  - Assumes same label distribution
    - Unlikely to be satisfied (especially in rare event scenarios)
- How do people make this problem easier?
  - Use estimated labels as opposed to actual labels
    - Klinkenberg, Ralf, and Thorsten Joachims. "Detecting Concept Drift with Support Vector Machines." *ICML*. 2000.

Detecting Concept Drift using a Classifier

- In $X \rightarrow Y$ problems, use labels generated by a classifier
  - Can not be used for $Y \rightarrow X$ problems

- Approach:
  - Estimate the next batch distribution using history of prior batch distributions
  - Pros: a good heuristic for jumps in concept
  - Cons: limited ability to monitor slow concept drifts
    - Requires significant change between batches
    - May not detect because concept drift affects classifier as well
      - Very likely since classifiers are trained based on labeled samples

Summary of Dataset Shift

- Dataset Shift detection has **NOTHING** to do with trained classifier performance.
  - Evidence that something about your training isn't consistent with testing.
  - A safe guard on the most fundamental assumption of machine learning (or any data-driven approach)
    - i.e., the test and training data are drawn from the same distribution.

- Dataset Shift: $p(x, y) \neq p'(x, y)$
  - Typically too hard to solve – generally regarded as impossible and avoided
    - There are likely exceptions, but those are tailored to specific scenarios

- Covariate Shift: $p(y|x) = p'(y|x)$ and $p(x) \neq p'(x)$ in $X \rightarrow Y$ problems
  - Lots of ways to detect effectively (see literature)
  - Consider hybrid (or ensemble) techniques to mitigate errors from a single test

- Prior Probability Shift: $p(x|y) = p'(x|y)$ and $p(y) \neq p'(y)$ in $\mathrm{Y} \rightarrow X$ problems
  - Can be solved using same techniques as Covariate shift
  - Requires either:
    - 1) ability to collect testing data labels
    - 2) assurance that changes in label distribution affect feature distribution
      - $p(y) \neq p'(y) \leftrightarrow p(x) \neq p'(x)$

- Concept Drift: Hardest of all dataset shift sub-problems
  - $p(y|x) \neq p'(y|x)$ and $p(x) = p'(x)$ in $X \rightarrow Y$ problem
  - $p(x|y) \neq p'(x|y)$ and $p(y) = p'(y)$ in $\mathrm{Y} \rightarrow X$ problem
  - Usually requires either testing data labels, or a classifier used to predict labels
    - Be careful when claiming to detect concept drift using predicted labels!!!

Outline

- Reading Material Recap
  - 2-3 minute impromptu overview
- What are Dataset Shifts?
  - Examples in medicine and anomaly detection
- Types of Dataset Shifts
  - Covariate shifts, prior probability shifts, concept drift
- Common causes of Dataset Shifts
  - Can these be used to improve detection?
- Common Assumptions in Dataset Shift Detection
  - What needs to be true to perform dataset shift detection?
- Research challenges in Dataset Shifts

Common Causes of Dataset Shift

- Now that we know how to classify Dataset Shift lets discuss causes – and how we might monitor the causes?

- Non-stationary environments
  - Flat tire on a car
  - People with different physiology

Utilize contextual information

- Data generation process
  - System faults
  - Missing data
  - User error

Utilize System Information

Common Assumptions when Detecting Dataset Shifts

- Detecting dataset shifts is still a statistical test
  - All tests (even nonparametric) are based on assumptions
- Common assumptions include:
  - Independent and identically distributed (i.i.d.)
  - Exchangeable data (time-series version of i.i.d.)
    - Given a sequential sample, any permutation of the sample is drawn from the same distribution
- No tests for these assumptions in autonomy scenarios
  - Must be an intrinsic property of the data generation process
  - To test would require the ability to "re-run" the scenario
- Additional assumptions are necessary for some tests
  - E.g., t-test, F-tests, etc.

How can we Facilitate Dataset Shift Detection?

- Understand the data generation processes
  - You need to understand causality of your data and features.
  - Exploit anything else that can be useful

- Constrain your feature space.
  - Don't include additional data just because it is available.
    - Leads to poor dataset shift performance = lots of false alarms!!!
  - Exploit natural invariants in the data provided by application context

- Sample data sufficiently to prevent concept drift
  - Active area of research devoted to monitoring concept drift

- Redundancy in testing
  - Rarely is one test always optimal, try multiple test as an esemble

Class Poll - Revisited

- We monitor shifts in datasets as a proxy for classifier performance?

- Which is more susceptible to test data not matching training data, SVMs or DNNs?
  - e.g., SVMs are simple linear classifiers, DNNs are more complex

- Is end-to-end control a good or bad idea? Why?
  - e.g., avoid feature engineering – just train a model to produce actuation commands.

Research Challenges in Dataset Shifts

- Develop monitors for specific data sources
  - The data from a camera/lidar is different than accelerometer
    - How can the specific data source influence dataset shift monitoring?
- Develop application-specific dataset shift monitors
  - e.g., develop a dataset shift monitor for an anomaly detector/medical alarm
- Explore worst-case and bounded dataset shift monitors
  - Many monitors are statistical in nature – what about applying techniques based on bounded errors?
- Investigate time-series dataset shift monitors in real systems
  - How do underlying dynamics affect time-series dataset shift detection?
- Learning in the presence of dataset shifts
  - Lots of work here already (e.g., active learning, etc.)
  - Could be interesting to apply to anomaly detection

THANK YOU!

PRECISE

PENN RESEARCH IN EMBEDDED COMPUTING AND INTEGRATED SYSTEMS ENGINEERING

http://precise.seas.upenn.edu