Dataset Shift

Sussex Data Analysis Forum

David Spence

9th November 2016

Me

- Yr 3 PhD
- Informatics
- ”Quantification under datset shift”
- d.spence@sussex.ac.uk
- Facilitator NOT expert!

Motivation

- You train a classifier/quantifier from the random 1% Twitter feed
- You apply it to estimate the gender balance of a group commenting on retirement homes in Scotland on Twitter

Common Terminology

- “Domain adaptation”
- “Fractures between data”
- “Concept shift”
- “Changing environments”
- “Population drift”

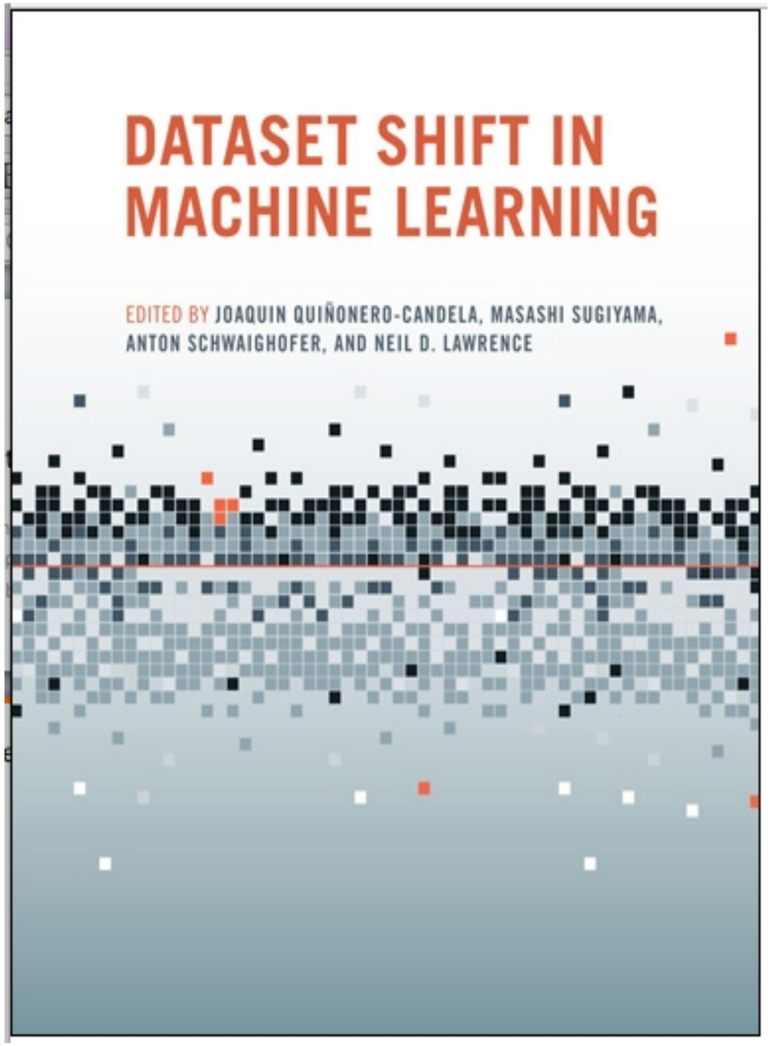
Dataset Shift

- "Most research in machine learning, both theoretical and empirical, assumes that models are trained and tested using data drawn from the same fixed distribution…In many practical cases, however, we wish to train a model in one or more *source* domains and then apply it to a different *target* domain"
- Shai et al. 2010

Dataset Shift

- “The most basic assumption used in statistical learning theory is that training data and test data are drawn from the same underlying distribution. Unfortunately, in many applications, the *in-domain* test data is drawn from a distribution that is related, but not identical to, the *out-of-domain* distribution of the training data”
- Daume III and Marcu, 2006

Dataset Shift


DATASET SHIFT IN MACHINE LEARNING

EDITED BY JOAQUIN QUIÑONERO-CANDELA, MASASHI SUGIYAMA, ANTON SCHWAIGHOFER, AND NEIL D. LAWRENCE

- Term “dataset shift” coined in this 2009 book
- Collection of papers from a NIPS workshop

Dataset Shift

- “…where the joint distribution of inputs and outputs differs between training and test stage”
- “…machine learning techniques assume that training and test distributions are identical”
- Quinonero-Candela et al. 2009

Typical Use Cases

- Spam detection
- Disease identification
- Credit approval
- Natural Language Processing

Key Sources

- Amos J Storkey, 2009
  - When Training and Test Sets are Different: Characteristic Learning Transfer
- Jose G. Moreno-Torres et al. 2012
  - A unifying view on dataset shift in classification
- Kull and Flach 2014
  - Patterns in dataset shift

Terminology

- X
  - Covariates
  - Features
  - Independent variables
  - Attributes
- Y
  - Target
  - Label
  - Dependent variable

Causal Direction

- Y → X
  - Features caused by class
  - E.g. disease (y) and symptoms (x)
- X → Y
  - Class caused by features
  - E.g. plays golf (y) and weather/schedule (x)
- Relevant to types of dataset shift
- See:
  - Webb and Ting (2005)
  - Fawcett and Flach (2005)

Joint Distribution

- $P(x,y)_{tr} \neq P(x,y)_{te}$
- So
- $P(x|y)_{tr}P(y)_{tr} \neq P(x|y)_{te}P(y)_{te}$
- And
- $P(y|x)_{tr}P(x)_{tr} \neq P(y|x)_{te}P(x)_{te}$

- Some assumption must be made about the relationship between the training and test data
- If we assume there is no relationship then the task is impossible!

Taxonomy

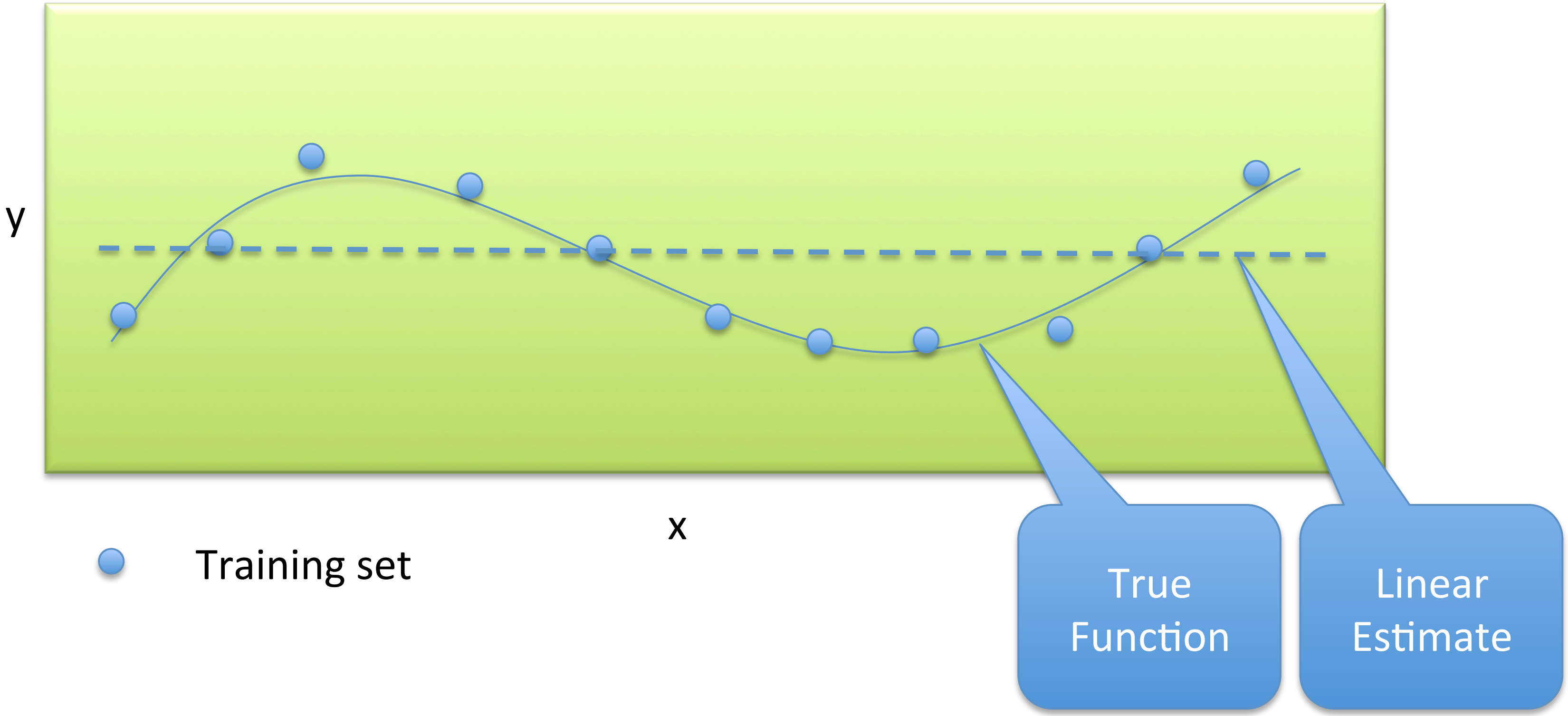
- Covariate Shift
- Prior Distribution Shift
- Concept shift
- Other…

Covariate Shift

- $P(y|x)_{tr} = P(y|x)_{te}$
- $P(x)_{tr} \neq P(x)_{te}$
- "Only in X → Y problems"
- Term coined by Shimodaira (2000)

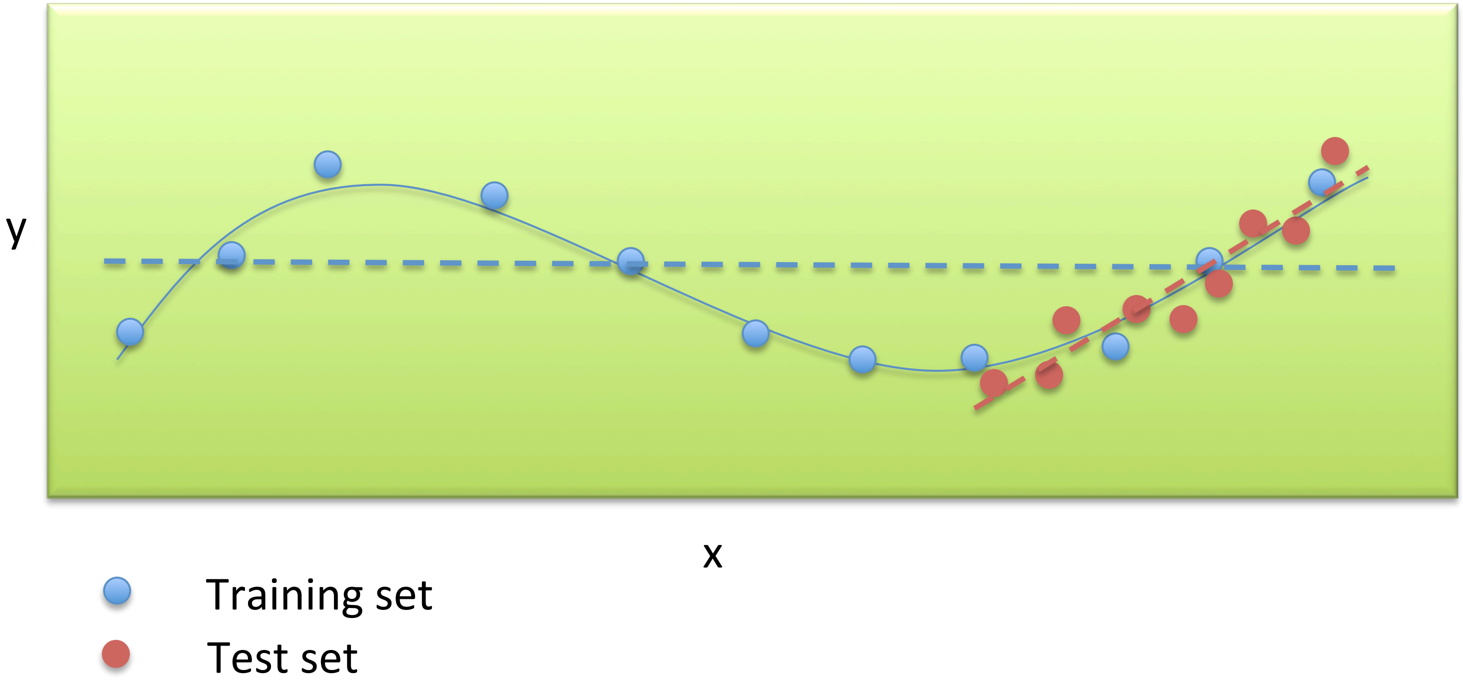
Why is this a problem?

- "…but P(y|x) doesn't change"



Why is this a problem?

- “…but $P(y|x)$ doesn’t change”

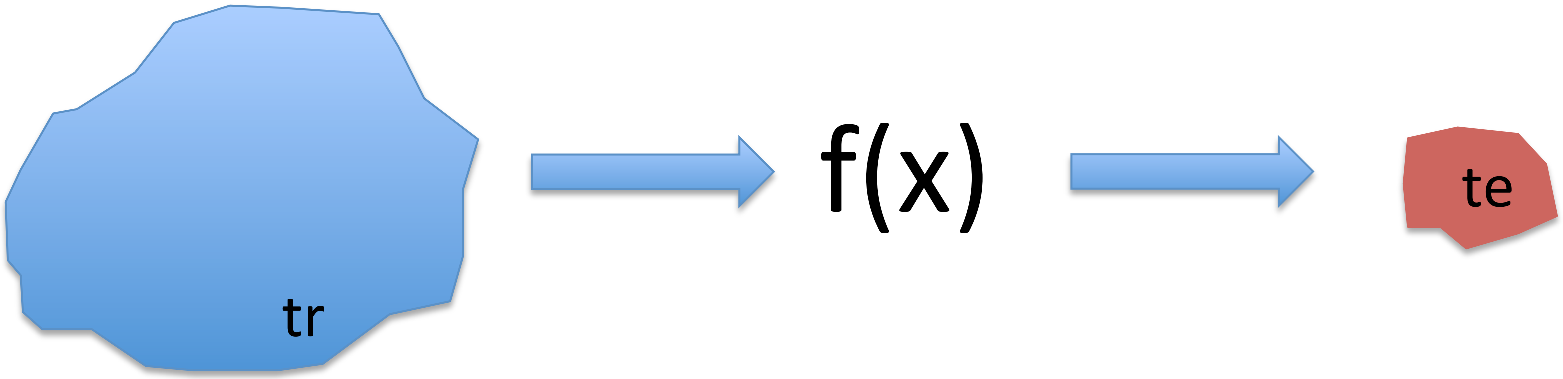


Mis-specified models
Potentially better fit to the true function for the area of interest

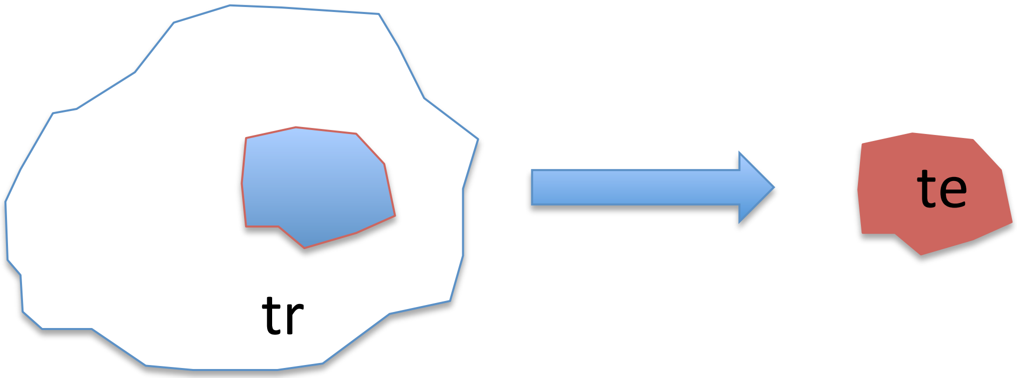
Transduction

- Vapnik 1998
- Transduction vs. Induction
- Do not solve a more general problem as an intermediate step
  - An estimated function for all the training data
- Try to get directly to the answer you actually want
  - The estimated classes of the test data

Induction



Transduction?



Approaches

- Reweighting training data to match the distribution of the test data
  - Minimising some “distance” between test and training domains
  - Kernel Mean Matching
  - Augmented feature space
  - Mapping features into a new representation
  - Loss function combining domain domain divergence and classification error
  - Structured Correspondence Learning
  - Lots of approaches…

Covariate Shift: References

- References
  - Shimodaira (2000)
  - Gretton, Smola et al (2009)
  - Daume III et al (2009)
  - Zhang et al (2013)
  - Etc.

Prior Probability Shift

- $P(x|y)_{tr} = P(x|y)_{te}$
- $P(y)_{tr} \neq P(y)_{te}$
- “Only in Y → X problems”

Approaches

- Reweighting training data
- Adapt classifier parameters according to the unlabelled test data
- See Saerens et al (2002)

Concept Shift

- $P(y|x)_{tr} \neq P(y|x)_{te}$
- OR
- $P(x|y)_{tr} \neq P(x|y)_{te}$

Approaches

- Try re-weighting approaches as for Covariate Shift and Prior Distribution Shift?
- Some assumption on the relationship between training and test data is still required
- Mixture of domains concept

Mixture Component Shift

- While
- $P(x,y)_{tr} \neq P(x,y)_{te}$
- Assume:
- $P(x,y|s)_{tr} = P(x,y|s)_{te}$
- For each sub-domain, s, and
- $P(s)_{tr} \neq P(s)_{te}$

Approaches

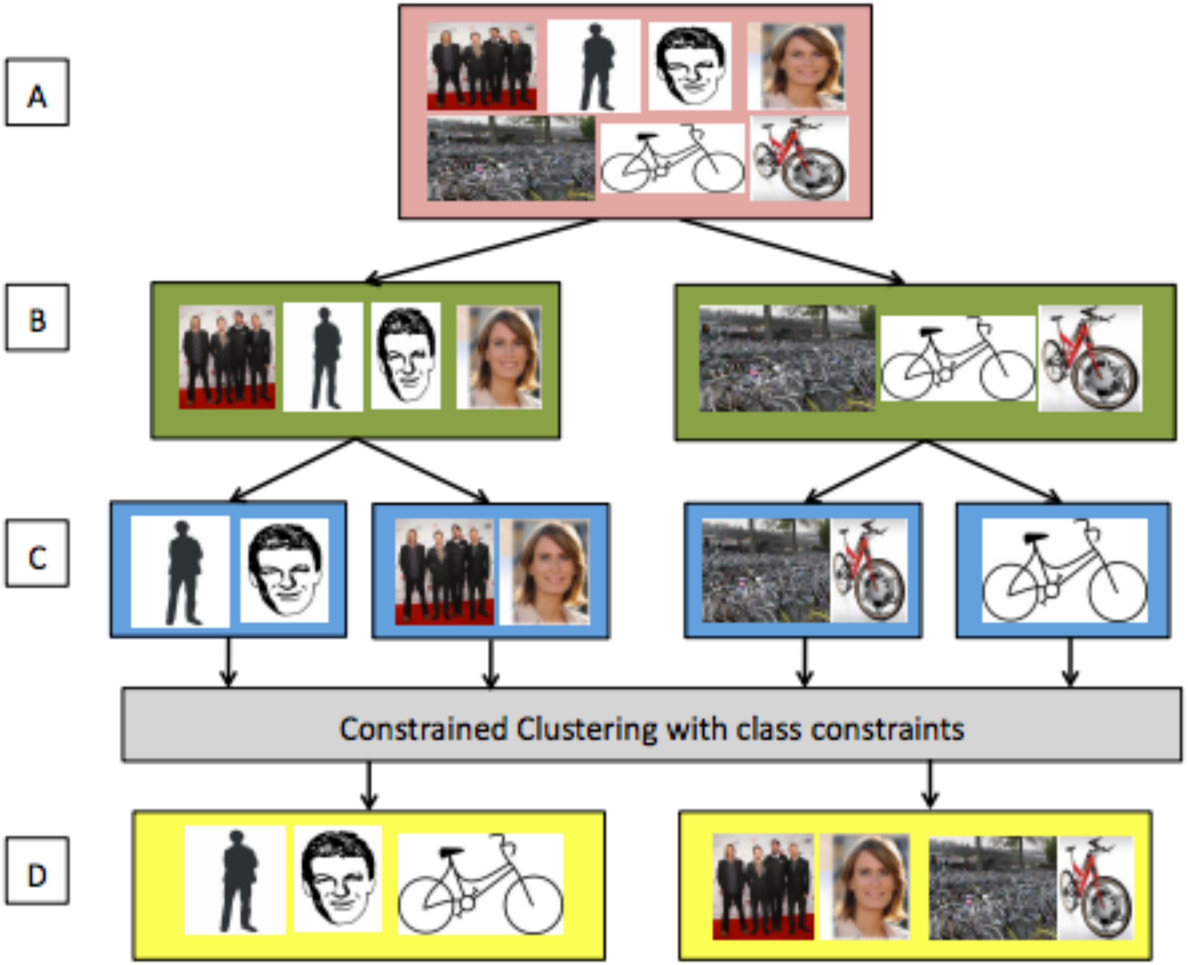
- Clustering within classes and aggregating clusters across classes
  - Hoffman et al (2012)

Hoffman et al (2012)

Labelled training set

Partition by class label

Cluster within class



Approaches

- Expectation maximisation algorithm for identifying implicit domains
  - Alaiz-Rodriguez et al (2011)

Causes of Dataset Shift

- Sample Selection Bias
- Missing At Random
- Missing Not At Random
- Etc.

- See Moreno-Torres (2012)