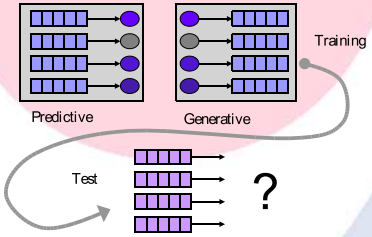
Mixture Regression for Covariate Shift

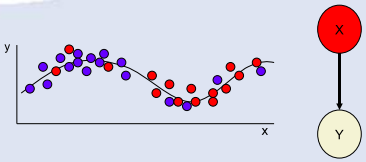
Amos J Storkey, School of Informatics, University of Edinburgh *a.storkey@ed.ac.uk*
Masashi Sugiyama, Department of Computer Science, Tokyo Institute of Technology *sugi@cs.titech.ac.jp*

ISSUE: What to do when training and test sets are *DIfferenT*.

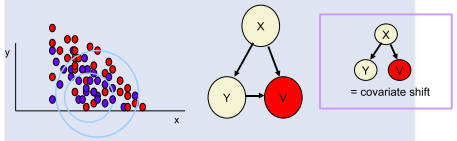


The situations where the training and test data are from different distributions is called *covariate shift*. There are a variety of causes of covariate shift:

Simple Covariate Shift: If we are correctly modelling the situation with a conditional model, and the only change is different covariate distributions, there are no real modelling issues.

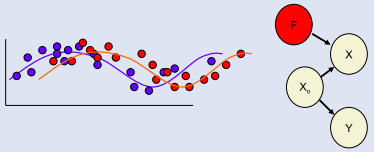


Sample Selection Bias: Sample selection rule (V in figure below) determines what samples occur in data. We need to estimate the sample rejection process.

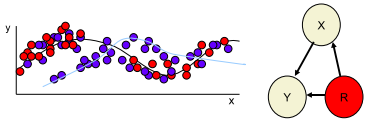


Domain Shift: The covariate X 'moves'.

$$X_{new} = f(X_{old}) \, , Y(X_{new}) = Y(f(X_{old})) \quad (1)$$



Source component shift: Proportions of different source components vary between datasets. Within source conditional models are same.



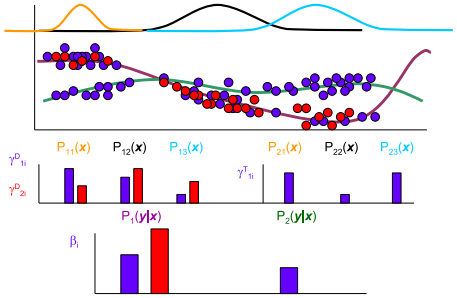
Mixture of Regressors for Source Component Shift

The model takes the following form

- The distribution of the training data and test data are denoted P_D and P_T respectively, and are unknown in general.
- Source set 1 consists of M mixture distributions, where mixture t is denoted $P_{1t}(\mathbf{x})$. Each of the components is associated with regression model $P_1(\mathbf{y}|\mathbf{x})$.
- Source set 2 consists of M_2 mixture distributions, where mixture t is denoted $P_{2t}(\mathbf{x})$. Each of the components is associated with the regression model $P_2(\mathbf{y}|\mathbf{x})$.
- The training and test data distributions take the following form:

$$P_D(\mathbf{x}) = \sum_t \beta_1 \gamma_{1t}^D P_{1t}(\mathbf{x}) + \beta_2 \gamma_{2t}^D P_{2t}(\mathbf{x})$$

$$P_T(\mathbf{x}) = \sum_t \gamma_{1t}^T P_{1t}(\mathbf{x})$$



EM Algorithm and Examples

This is a simple mixture of experts model, and it is straightforward to apply the EM algorithm for updates. We applied this for various regressors including Gaussian process regressors using a variational approximation for the mixture. We also show that Importance Weighted Least Squares is a special case.

Generated Test Data

We compare mixture of regressors approach to covariate shift (MRCS) with importance weighted least squares estimator (IWLS) given the best mixture model fit for the data and a mixture of regressors model that ignores the form of the test data, but chooses a regressor by matching to the test data distribution using a KL divergence measure (MRKL). The third case is where the mixture of regressors is used simply as a standard regression model, ignoring the possibility of covariate shift (MRREG). (100 datsets, linear, random numbers of mixtures, 8 restarts, 80 iterations of EM. Analysis was done for fixed model sizes and for model choice using a Bayesian Information Criterion (BIC).)

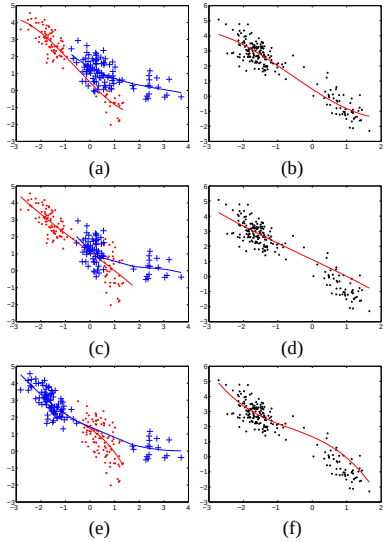
	MRCS	IWLS	MRKL	MRREG
1 Mixture	0.588	0.797	3.274	0.890
2 Mixtures	0.536	0.804	2.673	0.881
3 Mixtures	0.601	0.831	3.390	0.887
4 Mixtures	0.623	0.817	2.823	0.894
5 Mixtures	0.612	0.837	2.817	0.898
BIC Choice	0.6100	0.7990	2.8638	0.8813
MCRS better	-	77/100	72/100	84/100

Auto-mpg

To demonstrate covariate shift we can consider a reduced UCI Auto-mpg prediction task trained on cars from one place of origin and tested on cars from another place of origin. Here we consider predicting the fuel consumption (attribute 1) using the four continuous attributes. We train the model using data on cars from origin 1, and test on cars from origin 2 and origin 3. We use Gaussian process regressors for each regression function. The results of running this are in the table below

	GP	MRCS	IWLS	MRKL	MRREG
Origin 2	1.192	0.600	0.700	1.2243	0.7397
Origin 3	0.898	0.568	0.691	1.3862	0.706

The following figures show an example of the approach.



Nonlinear regression using covariate shift. (a),(c),(e) Training set fit and (b),(d),(f) test data with predictions for MRCS (top), IWLS (middle) and MRKL (bottom) respectively. In (a),(c),(e), the '.' data labels the points for which the test regressor has greater responsibility, and the '+' data labels points for which the training only regressor has greater responsibility.

Future Work

This framework is currently being extended to the case of multiple training and test datasets using a fully Bayesian scheme, and will be the subject of future work. In this setting we have a Topic model, similar to Latent Dirichlet Allocation, where each dataset is built from a number of contributing regression components, where each component is expressed in different proportions in each dataset. The model and tests of this paper show that this multiple dataset extension could well be fruitful.

School of informatics