Reproducing Sugiyama 2007

This is a basic exercise in covariate shift importance weighting. The simulation exercise described by Sugiyama *et al.* (2007) and a simple solution using weighted least squares is implemented here.

Importance Weighting

When performing maximum likelihood estimation, the goal is to maximize the expected value of a likelihood function $L(X, Y)$ over the joint distribution of feature-target pairs $P(Y, X)$. This is estimated as the sample average $\frac{1}{N}\sum_{i=1}^{N} L(X_i, Y_i)$.

Now consider a distribution of feature-target pairs in a source domain with a probability density function $P_S(X, Y)$ and a target domain with a probability density function $P_T(X, Y)$. One may have labelled feature-target pairs from the source distribution but not from the target distribution.

How could one train a model that maximizes the expected value of the likelihood $L(X, Y)$ over the target joint distribution? One could assume that $P_S(X, Y) = P(Y|X)P_S(X)$ and $P_T(X, Y) = P(Y|X)P_T(X)$, which means that the conditional distribution $P(Y|X)$ does not change between the two domains, and therefore only the distribution of the covariates changes. This is known as the covariate shift assumption. Assuming this, one could estimate the expected value of the likehood function as $\frac{1}{N}\sum_{i=1}^{N} \frac{P_T(X)}{P_S(X)} L(X_i, Y_i)$. However, this would require the densities of the unknown covariate distributions in the source and target domains.

Consider further a new combined distribution resulting from the mixture of both domains, with probability density function $P_C(X) = \gamma_S P_S(X) + \gamma_T P_T(X)$, where γ_S and γ_T are weights assigned to each domain in the mixture. This could be rewritten as

$$P_C(X) = P_C(S)P_C(X|S) + P_C(T)P_C(X|T),$$

where $P_C(T) = \gamma_T$ and $P_C(X|T) = P_T(X)$.

Given a sample from this mixture and domain labels, it's possible to train classifier to estimate the probability of the observation belonging to the target distribution, $P_C(T|X)$.

Further, we know that, by the Bayes theorem

$$\frac{P_C(T|X)P_C(X)}{P_C(S|X)P_C(X)} = \frac{P_C(X|T)P_C(T)}{P_C(X|S)P_C(S)},$$

and therefore

$$\frac{P_T(X)}{P_S(X)} = \frac{P_C(X|T)}{P_C(X|S)} = \frac{P_C(T|X)P_C(S)}{P_C(S|T)P_C(T)}.$$

This suggests that we could estimate the necessary importance weight as $\frac{g_C(T|X)N_S}{(1-g_C(T|X))N_T}$, where $g_C(T|X)$ is a an estimate of the conditional probability obtained by training a classifier from a sample of the combined domain having N_S observations from the source domain and N_T observations from the target domain. Note that no labels (Y) are needed in this process.

Outputs

Simple Illustrative Example: Extrapolating a linear model on the sinc function

As described by Sugiyama, we draw the training and test samples from two different gaussian distributions (the source domain and the target domain, respectively), as in the figure below:

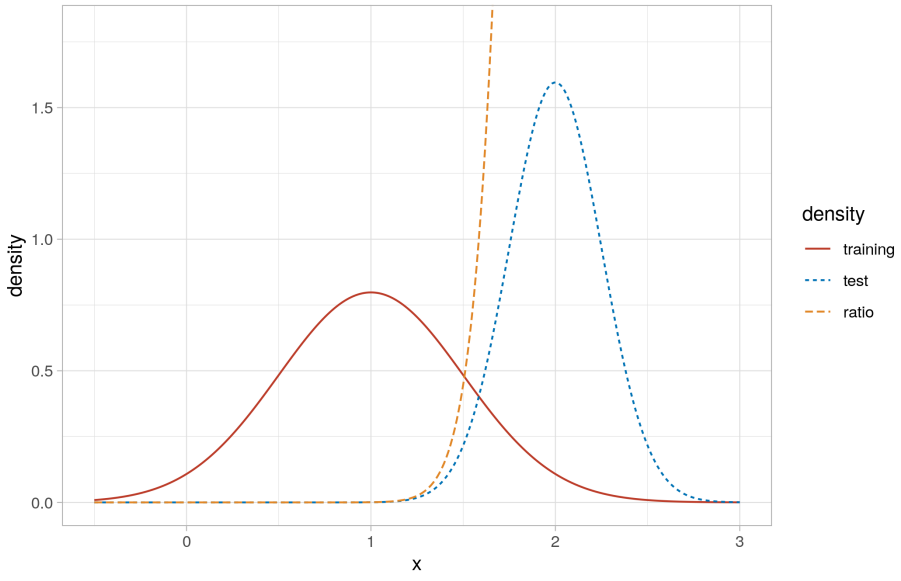

Figure 1: Training (Source) and Test (Target) Distributions, and Ratio of their Densities

An unweighted linear regression model is fit on the training data, which is unsurprisingly not accurate at predicting the test data, since the linear approximation is optimized locally on the region where the training data is sampled from (the source domain).

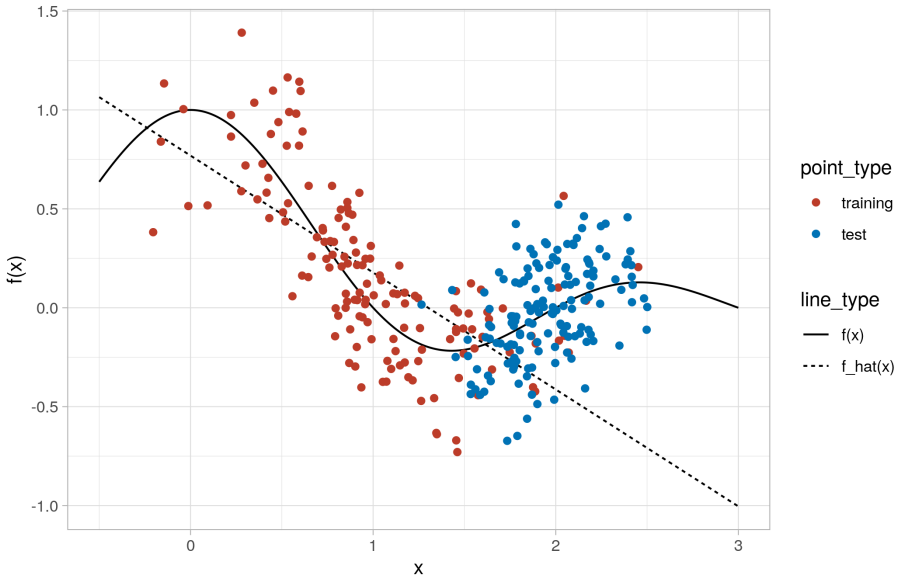

Figure 2: Unweighted Linear Regression

A weighted linear regression model is then fit, using a logistic regression classifier to estimate importance weights, as explained in the previous section. The model approximates the function in the region where the test data is sampled from (the target domain) much better.

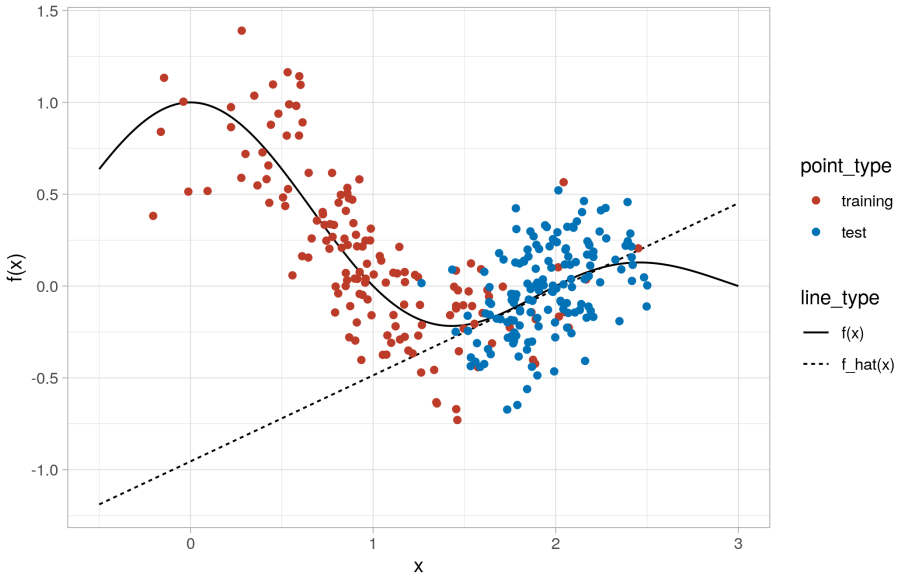

Figure 3: Weighted Linear Regression

In addition, we can see that if we train a model on a training data sampled from the target domain, the resulting linear approximation is very similar to that resulting from the weighted linear regression model.

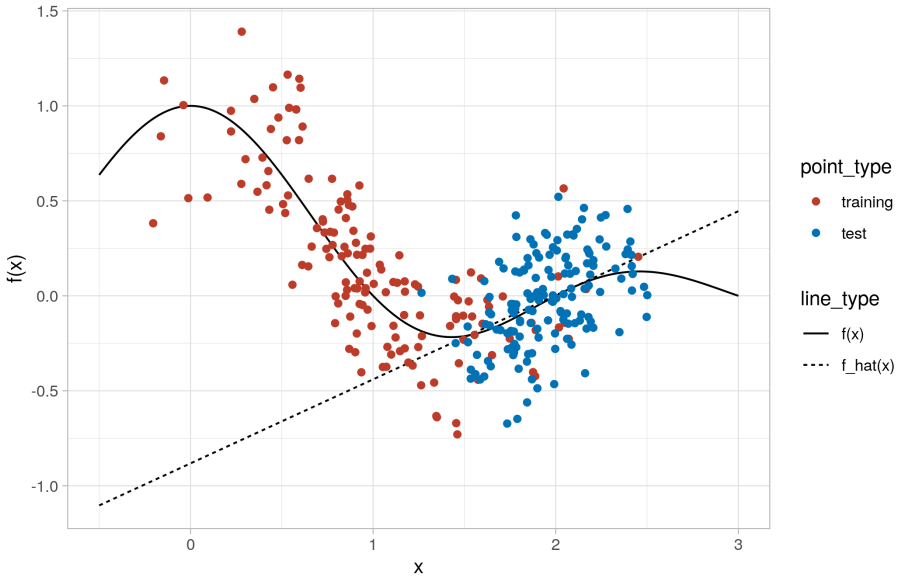

Figure 4: Linear Regression Fit on Target Domain

References

Sugiyama *et al.*, Journal of Machine Learning Research 8 (2007) 985-1005.