A Simple Example Framework for Assignment #2

This is a shorter version of what I expect, but the depth of investigation presented is what I expect.

Advanced topic area: Violation of the stationary distribution assumption.
Paper: Zadrozny

a) A one-sentence summary of the paper
b) An overview of the paper and its contributions.
c) Three open issues that the paper raises (in part 2 you will investigate this issue).

A1) Sample bias can be decomposed into four types, global learners are effected by sample bias, local learners are not and a method of correcting sample bias for global learners is presented.

A2)

There are four types of sample bias introduced:

1) A random “bias”
2) $P_{Train}(x) \neq P_{Test}(x)$, $P_{Train}(y) = P_{Test}(y)$, $P_{Train}(y|x) = P_{Test}(y|x)$
3) $P_{Train}(x) = P_{Test}(x)$, $P_{Train}(y) \neq P_{Test}(y)$, $P_{Train}(y|x) = P_{Test}(y|x)$
4) $P_{Train}(x,y|x_s) \neq P_{Test}(x,y|x_s)$, $P_{Train}(y) = P_{Test}(y)$, $P_{Train}(y|x) = P_{Test}(y|x)$

The third type of bias has been extensively addressed in the data mining community, the fourth type has been addressed in the econometrics community. The focus of the paper is the open issue of addressing the second type of sample bias.

The author decomposes learners/mining into those that are global and local.

She claims that local learners are not effected by sample bias and global learners are.

A3)

Issue #1: A significant issue is that global learners are affected by sample bias and local learners are not. However, the author has only experimentally verified the issue for one data set.

Issue #2: The claim that decision tree learners are global learners is not convincing. A local learner estimates $P(y|x)$ and so does a DT learner. The claim that generally $P(y|x) \neq P(y|x,s=1)$ is not backed up.
1. Discuss this item in greater detail than Part A) (1-2 pages).
2. Describe a principled approach to investigate the area (2 pages).
3. Describe in detail the methodology for your investigation. (2 pages).

Your aim is to convince the reader of two items:
   a. Your proposed area is non-trivial (for grads I expect one significant issue investigation).
   b. Your methodology will thoroughly investigate the proposed topic area.

We are focusing on:

“Issue #1: A significant issue is that global learners are affected by sample bias and local learners are not. However, the author has only experimentally verified the issue for one data set.”

B1)
We would like to extend the above issue with regard to the following points:

a) Further experimental results on other data sets to verify the authors claims for global versus local learners and learning from biased samples.

b) The author show that the models do indeed change. But typically in mining we are interested in accuracy (i.e. a 0-1 loss perspective). For instance, with naïve Bayes we know the modeling assumptions are violated (note the stationary distribution assumption is just that, an assumption) but we still get good predictive results. Can’t changing the marginal probability of an instance P(x) slightly not effect a learner’s actual performance, how big should the change be before the “accuracy” suffers.

B2)
For point a) we need only repeat her experiments but for different data sets. I will limit my experiments to only two learners: NB and DT.

For point b) We shall use the same data sets as point a). For each learner we will:

   a) take a training sample set of size n with alpha bias
   b) take a training sample set of size n with no bias
   c) an unbiased test set to score the learners performance
   d) For each test set instance report the difference in the predictions using the biased and unbiased samples
   e) Repeat a), b), c) five times
   f) Increase sample bias
For this section, you need to fill in B2)

a) take a training sample set of size n with alpha bias

We define alpha bias as being the average change in the marginal column probabilities is alpha. That is for a $m$ Boolean column training set then

$$\text{Sum}_{i=1}^{m} |P_{\text{Train}}(x_i) - P_{\text{Test}}(x_i)| \leq \alpha.$$ 

You would need to explain how you introduce this bias.

b) take a training sample set of size n with no bias
c) an unbiased test set to score the learners performance
d) For each test set instance report the difference in the predictions using the biased and unbiased samples

I will measure the average differences for naïve Bayes learners for the joint probability ($|P_{\text{Bias}}(y^*,x_{test}) - P_{\text{NoBias}}(y^*,x_{test})|$). For DT learners I will measure the average differences for ($|P_{\text{Bias}}(y^*|x_{test}) - P_{\text{NoBias}}(y^*|x_{test})|$) for $y^*$ is the true label.

In addition I will measure the average difference in the accuracy.

e) Repeat a), b), c) five times
f) Increase sample bias

C1)
The results are:

C2)
It appears that