Value Weighted Analysis: Building Prediction Models for Data with Observation Weights

December 12, 2001

1 Introduction

This paper discusses the analysis and modeling of data with observation weights. These *known* weights (which we will mostly refer to as observation values) represent how important each observation is, in terms of the size of population it represents, the amount of money it represents etc. Some examples of data with observation weights:

1. Monetary value: One example is credit scoring for loan approval, where each requested loan has to be labeled good/bad. The size of the requested loan is a good candidate for the value of an observation. Another monetary value example is attrition analysis in telecommunication, where the company is attempting to predict what customers will choose to terminate service with it. Valuable customers are obviously much more important, and so a customer's monetary value (such as the average bill size) may be a reasonable value indicator.

2. Sampling weights: Population samples often appear with sampling weights, for example in the case of census data, where each observation typically has a weight indicating the size of population it represents ([IMPACT 00]). Stratified samples are another example, and there is a rich literature on the proper use of sampling observation weights ([Korn and Graubard 95a], [Potter 90]).

3. Partial models for data: In some cases, we may be interested in prediction models for specific population segments (e.g. a specific city). In that case, we can argue that observations from this segment should have weight one, and all the rest weight zero. As we will see it may be advisable to bound the weights from below by some small $\epsilon > 0$, since we can then use these observations to reduce the variance of our predictions.

4. Numeric optimization: The classic example here is Boosting, in particular AdaBoost, where as ([FHT 01], chapter 11) have shown, weighted analysis is a tool for stepwise optimization of the underlying exponential criterion.

Most of the previous statistical work on use of observation values is concerned with sampling weights ([Korn and Graubard 95a], [Korn and Graubard 95b], [Potter 90], [Pfeffermann 93] and more). All of these concentrate on parameter estimation in a regression setup. They suggest various solutions for the bias-variance tradeoff involved in using the observation values in one of the two extreme approaches:

- *Ignore data values:* this approach advocates building a model while ignoring the values. This generally minimizes the variance of the parameter estimates (we prove an explicit result in section 3 below). However it may result in biased estimates (e.g. [Korn and Graubard 95a]).
- *Perform weighted analysis with the values as weights:* This approach assures us of unbiased estimates under reasonable conditions, however the variance of these estimates will generally be larger, much larger if the weight distribution is highly skewed.

In this paper we are interested in investigating useful ways of utilizing observation values for building prediction models. While survey methodology is concentrated around parameter estimation, the wider class of problems which we include under Value Weighted Analysis makes prediction a highly relevant methodological issue. We are also interested in considering a wider family of modeling techniques, not limiting ourselves to parametric regression models. This leads us in a different direction than the sampling literature, although some results from that work are useful for us, and we believe that our results have useful implications in the parameter estimation context.

We investigate the situations where each of the two approaches above would be optimal for prediction and argue that in many practical situations the best approach in terms of value weighted future performance of our model is actually a "middle ground" between the two extreme approaches. Namely, we advocate using a "toned down" version of the weights for model building. The extent of "toning down" can be approached as a tuning parameter and estimated from the data, e.g. by cross validation. One such approach - weight trimming - is discussed in the sampling literature ([Potter 90]). We prefer a different alternative - a power transformation - although most of our results are valid for both toning approaches.

The structure of this paper is as follows: Section 2 offers a mathematical formulation of the value weighted analysis (VWA) framework in the context of prediction models. A detailed theoretical discussion of VWA in linear models with squared error loss is offered in section 3. We show the bias-variance tradeoff involved in VWA, and give a simulation example. Section 4 discusses some other modeling procedures and loss functions and generalizes the results of section 3 to intuitions about VWA in general. In section 5 we implement the VWA framework on real datasets and show how the theoretical concepts discussed display themselves on real data. Section 6 reviews the literature on estimation for weighted sampling and relates it to our work and results. Some

new techniques for parameter estimation from weighted samples are suggested and discussed.

2 Value Weighted Analysis framework

The standard supervised learning framework is as follows:
We are given a set of training observations $\{x_i, y_i\}_{i=1}^n$, and a loss function $L(y, \hat{y})$. Our goal is to build a prediction model to minimize expected loss over *future, unseen* data. So we want to find a model: $\hat{y} = f(x)$ to minimize $E_{XY}L(Y, \hat{y}(X))$ over future data.

In the VWA framework we add one more ingredient to the mix. Each observation has a "weight" or "value" attached to it. That is, our data is now $\{x_i, v_i, y_i\}_{i=1}^n$ where v_i is the value of the ith observation. We are going to focus exclusively on situations where the loss function is linearly value-weighted, i.e. can be expressed as $L_v(y, \hat{y}) = v \cdot L(y, \hat{y})$. Examples:

1. Misclassification rate:
   - Non-weighted goal: $minP_{XY}(Y \neq \hat{y}(X))$
   - Value-weighted goal: $minE_{XYV}V \cdot 1\{Y \neq \hat{y}(X)\}$
2. Squared loss:
   - Non-weighted goal: $minE_{XY}(Y - \hat{y}(X))^2$
   - Value-weighted goal: $minE_{XYV}V \cdot (Y - \hat{y}(X))^2$

We assume that the values are "like" the predictors, *not* part of the response. Hence when we have to predict future data we will know its values, and the value feature can even serve as an additional predictor.

We want to investigate correct use of the *training set* values when building a prediction model. So assume we have some black box model-generator M which builds a model by attempting to minimize some loss function L^* (which may or may not be the same as L) over the training data. Given training data $(X, y) = \{x_i, y_i\}_{i=1}^n$, we can generate a model:

$$\hat{m} = M(X, y) \approx \arg\min_m \sum_{i=1}^n L^*(y_i, m(x_i)) \tag{1}$$

Where we have $\approx$ because M may optimize over a limited family of models or may not be able to find the global optimum. For example, if M is linear regression and L, L^* are squared loss, then we have that $\hat{m}$ is the least squares model, fitting:

$$\hat{m}(newx) = newx \cdot (X'X)^{-1}X'y \tag{2}$$

Now assume that we can also input weights $w = \{w_i\}_{i=1}^n$ into M, so we will get a weighted solution:

$$\hat{m} = M(X, y, w) \approx \arg\min_m \sum_{i=1}^n w_i \cdot L^*(y_i, m(x_i)) \tag{3}$$

We want to investigate the effect of initializing w in various ways given the observation values v. The two extreme situations mentioned above are:

- Ignore values: $\forall i,\ w_i = 1$
- Use values as is: $\forall i,\ w_i = v_i$

We now define a family of candidate transformations which we will use in the next sections. Given $0 \leq p \leq 1$, let the p-transformation of the values into weights be: $\forall i,\ w_i = v_i^p$. Of course, $p = 0$ amounts to ignoring values, while $p = 1$ means we are taking the values as-is. Intermediate powers would correspond to a compromise between these two approaches. The choice of the power transformation family is based on the idea that the transformation should bring strongly skewed weights towards equal weight gradually as p decreases. There are two other obvious candidates for a family of transformations:

1. Linear interpolation between $w_i \equiv 1$ and $w_i \equiv v_i$. This would not give us the desired effect, as the extreme values would remain extreme until p gets very close to 0.
2. Value trimming: this is the solution usually advocated in the survey literature. In this context, trimming means setting a threshold t and transforming: $v_i \rightarrow \min(v_i, t)$ (this is sometimes called "Windsorising" in the literatue). [Potter 90] reviews various approaches to choosing a trimming threshold. Treating the threshold as a tuning parameter negates the need to choose among these approaches. However this approach seems overly aggressive in its treatment of the most valuable observations. We believe that the power transformation approach may be superior both in the prediction and parameter estimation domains (see [Korn and Graubard 95a], section 3.4, for a sense in which superiority can be defined, as minimizing variance while remaining stochastically "unbiased"). However this is a subject of future research. In fact, *all* the theoretical results of sections $3.1 - 3.3$ and 4.4 hold for trimming as well.

2.1 Difference between VWA and Variance Weighted analysis

Weighted least squares is often used in cases where data is heteroskedastic (e.g [Weisberg] chapter ...). For example, if in homoskedastic data we have k_i independent observations at each point x_i we can combine them to one observation with weight k_i (since their mean $\bar{y}_i$ has variance σ^2/k_i). In that case, fitting the weighted model is equivalent to fitting OLS without combining the observations. In the more general case, variance-weighted least squares fits the maximum normal likelihood model given the covariance matrix for the y_i's. Weighted model fitting is simply a computational tool to achieve this goal.

By contrast, in Value-Weighted analysis, the weights do not correspond to variances, but rather to "importance" measures. Determining the weighting

scheme for model fitting is a modeling decision, with no direct likelihood interpretation (although a "pseudo likelihood" interpretation for sampling weights has been suggested - see [Pfeffermann 93], section 7 and [Pfeffermann et al. 98]).

3 VWA in linear regression models with squared error loss

Assume we have data $\{x_i, v_i, y_i\}_{i=1}^n$, and we are interested in minimizing the VW squared EPE (Expected Prediction Error) on a "future" copy of our data $\{x_i, v_i, \tilde{y}_i\}_{i=1}^n$. That is, we are assuming that the x's and v's are the same and that the y's are a new, independent copy. A more realistic assumption would be that the v's also change stochastically on the future data. However, the "constant v" assumption will allow us to formulate and prove general results. At the end of this section we discuss the implications of relaxing the assumption and conclude that the intuitions we gain from this analysis remain essentially unchanged. Our experiments in section 5 verify this. Throughout this section we also assume the response to be independent homoskedastic, in other words:

$$\forall i,\ var(y_i) = var(\tilde{y}_i) = \sigma^2,\ cov(y_i, \tilde{y}_i) = 0$$
$$\forall i \neq j,\ cov(y_i, y_j) = cov(y_i, \tilde{y}_j) = cov(\tilde{y}_i, \tilde{y}_j) = 0$$

So we are trying to find:

$$\beta^* = \arg\min_\beta EPE = \arg\min_\beta E_{\tilde{y}}[\sum_{i=1}^n v_i(\tilde{y}_i - \beta' x_i)^2] \tag{4}$$

How should we use the weights for calculating β^* to minimize VW EPE?

The key is the Bias-Variance decomposition for the EPE:

$$E_{\tilde{y}} \sum_{i=1}^n v_i(\tilde{y}_i - \hat{\beta}' x_i)^2 = \sigma^2 \sum_{i=1}^n v_i + \sum_{i=1}^n v_i var(\hat{\beta}' x_i) + \sum_{i=1}^n v_i (E\hat{\beta}' x_i - Ey_i)^2 \tag{5}$$

With the first component denoting the irreducible error, the second representing the weighted variance resulting from model building and the third representing the weighted squared bias.

For the standard (non-weighted) problem, we know that if the linear model is unbiased, i.e. if $Ey = X \cdot \beta$, then the least squares solution $\hat{\beta} = (X'X)^{-1} \cdot X' \cdot y$ is:

- Unbiased: $E\hat{\beta} = (X'X)^{-1} \cdot X'X \cdot \beta = \beta$
- Has smallest variance among linear unbiased estimators of β - this is the Gauss-Markoff theorem, i.e. if $\tilde{\beta} = Sy$ then $\forall x, var(x'\hat{\beta}) \leq var(x'\tilde{\beta})$.

In the next subsections we attempt to generalize this kind of optimality result to VWA with weighted least squares. We show that the non-weighted LS

solution still minimizes the variance uniformly among the family of p-weighted fits even if the model is biased. We then show that the model with $p = 1$ minimizes VW squared bias. This leads us to a bias-variance tradeoff view of weighting in linear model fitting.

3.1 Non-weighted LS solution minimizes variance

Let us denote by $\hat{\beta}_p$ the least squares solution using weights $w_i = v_i^p$, in other words:

$$\hat{\beta}_p = \arg\min_{\beta} \sum_{i=1}^{n} v_i^p \cdot (y_i - x_i' \cdot \beta)^2 \tag{6}$$

Proposition 1 *The variance of the fit is minimized when $p = 0$, in other words:*

$$\forall p, x, \;\; Var(x'\hat{\beta}_0) \leq Var(x'\hat{\beta}_p) \tag{7}$$

Proof: Let $W_p = diag(v_1^p, \ldots, v_n^p)$ then it is easy to verify:

$$\begin{aligned} \hat{\beta}_p &= (X'W_pX)^{-1} \cdot X' \cdot W_p \cdot y \\ Var(\hat{\beta}_p) &= \sigma^2 \cdot (X'W_pX)^{-1} \cdot X' \cdot W_p^2 \cdot X \cdot (X'W_pX)^{-1} \end{aligned}$$

So the variance does not depend on the bias of the model at all.
From the Gauss-Markoff theorem we know that for an unbiased model (i.e. if $Ey = X \cdot \beta$), we have $Var(x'\hat{\beta}_0) \leq Var(x'\hat{\beta}_p)$. But those variances are independent of the bias, so this fact holds for a biased model as well. □

Note, that this result only shows that $p = 0$ gives minimal variance predictions among *weighted least squares* fits. It is trivially *not* true that this model has minimal variance predictions among all models (note that we are *not* assuming unbiasedness), since the predictions of the linear model with $\hat{beta} \equiv 0$ have variance 0.

Corollary 1 *If the linear model is unbiased, then the unweighted solution $\hat{\beta}_0$ minimizes EPE among all weighted least squares solutions.*

Proof: Trivial consequence of proposition 1, since if the bias is 0, we only have to minimize variance. □

These results are pretty intuitive, since taking a power $p \neq 0$ is effectively reducing the amount of information we are using for determining the model parameters. This intuition should be useful for non-linear models and non-squared loss setups, where we cannot formulate variance results formally.

3.2 The LS solution with $p = 1$ minimizes VW squared bias

Proposition 2

$$\forall p, \;\; \sum_{i=1}^{n} v_i \cdot [E(y_i) - x_i' \cdot E(\hat{\beta}_1)]^2 \leq \sum_{i=1}^{n} v_i \cdot [E(y_i) - x_i' E(\hat{\beta}_p)]^2$$

Proof: Let $V = diag(v_1, \ldots, v_n)$ and define a V-inner product in $\Re^n$ as:

$$< z_1, z_2 >_V = z_1' V z_2 \tag{8}$$

Then if we set $\hat{y} = X\beta_1 = X(X'VX)^{-1}X'Vy$ we get:

$$\begin{aligned} < Ey - E(\hat{y}), E(\hat{y}) >_V &= (Ey)'(VX(X'VX)^{-1} \cdot X' - I) \cdot V \cdot \\ &\quad (X \cdot (X'VX)^{-1}X'V)(Ey) = \\ &= (Ey)'(VX(X'VX)^{-1} \cdot X'VX \cdot (X'VX)^{-1} \cdot X'V - \\ &\quad -VX \cdot (X'VX)^{-1} \cdot X'V)(Ey) = \\ &= 0 \\ \Rightarrow E(y - \hat{y}) \perp_V E(\hat{y}) & \end{aligned} \tag{9}$$

In other words, $E(\hat{y})$ is the V-projection of Ey, therefore it is the V-closest point to it in the linear subspace spanned by the X matrix.
This actually proves more than our claim - that $X\beta_1$ has the smallest VW square bias of *any* estimator that is linear in X. □

Propositions 1 and 2 combined give us some intuition about the bias-variance tradeoff involved in VWA. We see that the variance is uniformly minimized when we take $p = 0$, while the value-weighted bias is minimized when we take $p = 1$. The optimal power p^* will thus be determined by the balance between those two effects. In general, if the reducible error is dominated by variance (in particular if the model is unbiased), we can expect to get $p^* \approx 0$, while if the VW bias effect is much bigger then we can expect to get $p^* \approx 1$.

It is important to note, though, that the bias results were based on the notion that the values remain the same on the "future" copy of the data. This assumption strongly biases our results towards favoring value weighting. For example, imagine that in reality the future $\tilde{v}_i$ are all $i.i.d$ from some V-distribution, independent of the values of $(x_i, \tilde{y}_i)$. In that case, proposition 2 is completely irrelevant (since we cannot infer any useful information from the training V matrix) and we can easily show that we can do no better than the LS solution with $p = 0$ in terms of VW EPE (proposition 3 below proves a similar result). However, if the future $\tilde{v}_i$ are strongly correlated with the training set values, proposition 2 still has merit as an indicator that use of the training set values is likely to decrease future VW bias.

3.3 VW linear regression simulated example

We illustrate our results with a simple example. We have the following setup:

$$\begin{aligned} x_i &\sim U(0, 10), \quad i \in \{1, 2, 3\} \\ y &= x_1^2 + x_2^2 + x_3^2 + N(0, 38^2) \\ v &= (x_2 - 5)^6 + 1 \end{aligned}$$

We are building a linear model: $\hat{y} = \beta_{p,1} \cdot x_1 + \beta_{p,2} \cdot x_2 + \beta_{p,3} \cdot x_3$, where β_p is the weighted least squares solution with weights v_i^p. So the model is obviously

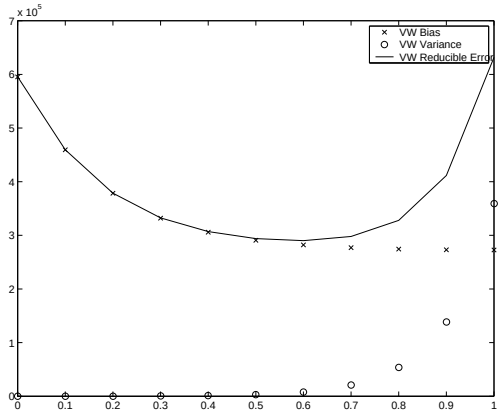

Figure 1: VW squared bias, variance and reducible squared error as a function of the power p

badly biased and we would expect to get an optimal power that is distinctly different from 0 and 1.

Figure 1 shows the VW variance, bias and reducible VW future prediction squared error as a function of the power p. As expected, we see that the variance is minimized at $p = 0$ and the VW bias is minimized at $p = 1$. The minimal VW reducible error is obtained at $p^* \approx 0.6$ and gives reducible error that is about *half* of the reducible error at either $p = 0$ or $p = 1$.

So we see in this example that the bias-variance tradeoff strongly favors an intermediate power over both of the "obvious" choices.

3.4 Viewing p as a regularizing parameter

As we have seen above, setting $p = 0$ amounts to minimizing variance. Setting $p = 1$ amounts to optimizing, on the training data, the same loss function whose expected future value we want to minimize (that is, we take $L = L^*$ in the notation of section 2).

So in the spirit of (...) we can view p as a penalizing or regularizing parameter, determining to what extent we are "smoothing" or reducing variance, relative to the most greedy fit. This can give us another view of the way in which taking $p << 1$ can serve us to build better, less variable models.

The regularization in this form is limited to what we can achieve by setting $p = 0$. However, we can also regularize independently with an additional penalty - Ridge penalty (...), Lasso (...) and others can all be used. It should be an interesting research question to inspect the effect that p-regularization has compared to those standard penalties in VWA.

3.5 Conditions for intermediate power to do well

Given our results from the previous sections we can try and formulate an intuitive description of when we expect intermediate powers to do much better than both $p = 0$ and $p = 1$.

We have already mentioned that if our reducible error is dominated by variance we would expect $p^* \approx 0$, while if it is dominated by VW squared bias *and* the value function is locally smooth and consistent between the training data and future data, we would expect $p^* \approx 1$.

For intermediate powers to do much better than both $p = 0$ and $p = 1$, we therefore need:

(TO BE CONTINUED...)

4 VWA for other loss functions and modeling approaches

The detailed discussion of linear models with squared error loss has given us a good intuition about VWA and the bias-variance tradeoff involved in it. We now turn to some other modeling approaches to see which of those results generalize and to what extent.

We discuss nearest neighbors models, decision tree methods and logistic regression, describe how VW analysis would be implemented in each case and consider its relevance. In section 4.4 we turn away from the discussion of specific modeling techniques and prove a general population result about cases where non-weighted analysis is certain to do better than VW analysis.

4.1 k-nearest neighbors models

Nearest neighbors models can be used for modeling either continuous or factorial responses. We concentrate on the continuous case, and assume as before that the response values y_i for both the training set and the future population are independent and homoskedastic with variance σ^2 . In our previous notation, given a new observation at point x we will predict it as:

- Non weighted: $\hat{y} = \frac{1}{k} \cdot \sum_{x_i \in N_k(x)} y_i$
- Weighted: $\hat{y}_v = \frac{1}{\sum_{x_i \in N_k(x)} v_i^p} \cdot \sum_{x_i \in N_k(x)} v_i^p \cdot y_i$

Where $N_k(x)$ is the the smallest neighborhood of x containing k training examples.

Let us consider intuitively the effect of value weighting on our prediction. To make it easier, we start by making a simplifying assumption, namely:

$$\forall x, \quad \sum_{x_i \in N_k} v_i = c_1$$

$$\sum_{x_i \in N_k} v_i^2 = c_2$$

So all neighborhoods are assumed to look the same in terms of value distribution.

Let us define $l = \frac{c_1^2}{c_2}$, the "equivalent" number of nearest neighbors ($l \leq k$, of course, by the Schwartz inequality), and consider the VW k-NN model vs. the non-VW l-NN model. Given a point x to predict at let $\hat{y}(x)$ be the VW k-NN prediction and let $\tilde{y}(x)$ be the non-VW l-NN prediction, then:

$$var(\tilde{y}(x)) = var(\hat{y}(x)) = \sigma^2/l$$
$$E(\tilde{y}(x)) = \frac{1}{l} \cdot \sum_{x_i \in N_l(x)} E(y_i)$$
$$E(\hat{y}(x)) = \frac{1}{c_1} \cdot \sum_{x_i \in N_k(x)} v_i E(y_i)$$

So the prediction $\tilde{y}(x)$ averages closer points to x, and since the NN underlying assumption is that choosing fewer (i.e. nearer) neighbors corresponds to reducing bias, we can generally expect the l-NN model to have smaller bias.

Overall, we get that the l-NN model predictions will have the same variance and smaller bias than the VW k-NN model predictions. If we now remove the simplifying assumption we made above, we get that for each fixed x the result still holds, however l will now depend on x, and so we cannot globally compare l-NN to VW k-NN. However we do get that at every x we can find a better non-VW nearest neighbors model. The intuitive lesson is that value weighting is unlikely to do any good in nearest neighbor models (as well as other local kernel models).

4.2 Decision trees

As a major tool in data mining and statistical learning, decision trees (e.g. CART - [BFOS 84] ,C4.5 - [Quinlan 93]) are an interesting modeling tool to investigate in the context of VWA. It should be noted that weighted analysis is implemented both in CART and C4.5. It is used by C4.5 for handling missing predictor values and by CART for ???. Decision tree algorithms have to make two main kinds of modeling decisions:

- Splitting decisions: these include whether to split, on what variable and at what value.
- Fitting decisions: given a tree, a fitted value (or class) has to be determined at every terminal node.

If we consider the second kind of fitting only (i.e. given a tree structure find the best fitting constants or classes), then we are essentially back to fitting a linear model at each terminal node. We can then refer back to the analysis from section 3 above.

The splitting decisions, on the other hand, are greedy steps based on splitting criteria (information gain for C4.5, Gini index or squared loss for CART), which are proxies for the true goal of minimizing misclassification rate (although the CART criterion for continuous response is actually the same for splitting and fitting). The greediness of the steps implies that the only statements we can make about whether or not weighting the splitting criteria would be advisable are intuitive ones. Since in the early splits in a decision tree we generally have a lot of data upon which to base the splitting decision, it seems reasonable that variance would be of little consequence and we would profit from value-weighting the splitting criteria. Once the data becomes more sparse as we advance down the tree, the bias-variance tradeoff may kick in. It seems that the best way to resolve this issue would be by doing data experiments.

4.3 Value weighted logistic regression

In the case when $y \in \{0, 1\}$, it is a popular approach to build a linear model for $logit(p) = \log \frac{P(y=1)}{P(y=0)}$. For an exposition of the fitting procedures see chapter ... of [McCullagh and Nelder 89] .

Instead of minimizing training misclassification rate (L in section 2 notation), logistic regression aims to maximize training set log-likelihood (L^*). By analogy to the linear regression case, we can integrate weights into the likelihood function and seek to maximize:

$$\sum_{i=1}^{n} v_i^p \cdot (y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \tag{10}$$

This formulation is also consistent with the "pseudo likelihood" view mentioned in [Pfeffermann 93], section 7, since if we had v_i^p observations at x_i this would be the likelihood function.

Recall the Fisher Scoring iterative algorithm for fitting a logistic regression model. Each iteration is a weighted least squares fit:

$$\begin{aligned} \beta_{k+1} &= (X'WX)^{-1}XWz_k \\ W &= diag(\hat{p}_{k,i} \cdot (1 - \hat{p}_{k,i})) \end{aligned}$$

Where z_k is a modified response.

It can easily be seen that VW Fisher scoring iterations would only entail integrating the likelihood weights into the weight matrix:

$$W \to W * diag(v_1^p, \ldots, v_n^p) \tag{11}$$

In section 5 we will see that applying this this approach to real data gives an empirical "Bias-variance" tradeoff similar to the one we have seen for linear models.

4.4 A general theorem for population optimality of unweighted model

We now switch to considering an arbitrary family of models $\mathcal{M}$ and loss function L. We are seeking a characterization of the best model $m^* \in \mathcal{M}$ minimizing $E_{XYV} V \cdot L(Y, m(X))$. If we assume that V is stochastically independent of the pair (X, Y), then we get the following simple characterization, implying that the best model for non-VW loss is also the best one for VW loss.

Proposition 3 *If the distribution of the values is independent of the distribution of both the dependent variable and all the other predictors, then the optimal model out of a given family $\mathcal{M}$ for non-weighted loss will be optimal in terms of future VW loss as well.*

Proof: Let $\tilde{m} \in \mathcal{M}$ be the model minimizing non-VW loss. Since V carries no information about Y, we can assume that the optimal model does not use V for prediction, i.e. $\tilde{m} := \tilde{m}(X)$, since else we can improve by taking:
$\forall x,\ \tilde{\tilde{m}}(x) = \arg\min_V E(L(Y, \tilde{m}(x, v)))$.
Then we get for some $m \in \mathcal{M}$:

$$E_{XYV} V \cdot L(Y, m(X, V)) = E_V E(V \cdot L(Y, m(X, V))|V)$$

Now, for $V = v$ we get:

$$E_{XY}(v \cdot L(Y, m(X, v))) \geq E(v \cdot L(Y, \tilde{m}(X)))$$

and thus (integrating over the distribution of V):

$$E_{XYV} V \cdot L(Y, m(X, V)) \geq EV \cdot EL(Y, \tilde{m}(X))$$

□

In other words, we have to minimize the non-weighted loss, to get an optimal model for future data. Since this is a population result without reference to sample variance or optimization considerations it cannot serve us directly in learning about fitting. However it gives us an intuition that if such independence indeed exists we are most unlikely to gain from using the values in any way for modeling. We will see below a real data example where this seems to be the case to some extent.

For the case of linear models with squared loss, we have seen in corollary 1 that unbiasedness assures optimality of non-weighted solution. Combining it with proposition 3 it is easy to see that the weaker condition of independence of bias and value is enough to assure optimality of the non-weighted solution in that case.

5 Real data experiments

We describe our experiments with two datasets from the UCI repository of Machine Learning and KDD datasets ([Blake and Merz 98], [Bay 99]). These

two examples illustrate the wide and varied range of problems which can be naturally interpreted as Value-Weighted prediction problems.

In both cases we have a binary dependent variable, and we are trying to minimize future VW misclassification rate. However in the first case, of loan data, the value indicator has to do with monetary gain/loss, while in the second case, census data, we have the classical weighted sampling situation.

We are using, as before, the parameter p as the power to which we exponentiate the training observation values to use as weights in building our models. In section 4.3 we have described how VW logistic regression can be implemented. We use this approach for our experiments.

5.1 German Credit Scoring data

Available from [Blake and Merz 98], under *statlog/german*:
ftp://ftp.ics.uci.edu/pub/machine-learning-databases/statlog/german/.

This dataset contains 1000 records of "good" and "bad" loans granted to customers of a German financial institution. The size of the requested loan is given, and this is a natural value indicator for our purposes, since both the loss from refusing a "good" loan and from granting a "bad" loan are proportional to the size of the loan. The original problem has an additional cost matrix for the different kinds of errors, which we ignore here although there is no problem in integrating it into VW analysis the same way we would into a non-VW one.

For the purpose of our experiments we used only the nine numeric predictors in the data. We built logistic regression models and varied the number of predictors used and the amount of data used for training (the rest was used as a test set).

As we have seen in the squared error case, the optimal power is a bias-variance issue. In the case of VW misclassification error, we cannot directly refer to these concepts. Many different bias-variance decompositions have been suggested for misclassification rate ([Friedman 97], [Breiman 96] are two examples). For our purposes, it suffices to consider the abstract idea:
- if reducible prediction error is governed by model uncertainty, then a lower power (value of p^*) would be optimal, since it allows more effective use of the data
- if reducible prediction error is dominated by the model's inadequacy in describing the goal distribution, then a higher power would be optimal, since it allows the model to concentrate on "valuable" regions.

Figure 2 shows three graphs of future VW misclassification vs. power used in training. Each graph displays the mean test set VW misclassification averaged over 5000 logistic regression models. The models were generated by randomly dividing the data into training and test sets. 95% confidence intervals (calculated empirically from the 5000 repetitions) are also displayed. As expected, increasing the number of predictors used decreases the optimal power, since it causes more model uncertainty and decreases the systematic error ("bias"). Thus controlling variance becomes more critical and the optimal power is lower.

Increasing the size of the training data increases the optimal power, since it decreases model uncertainty.

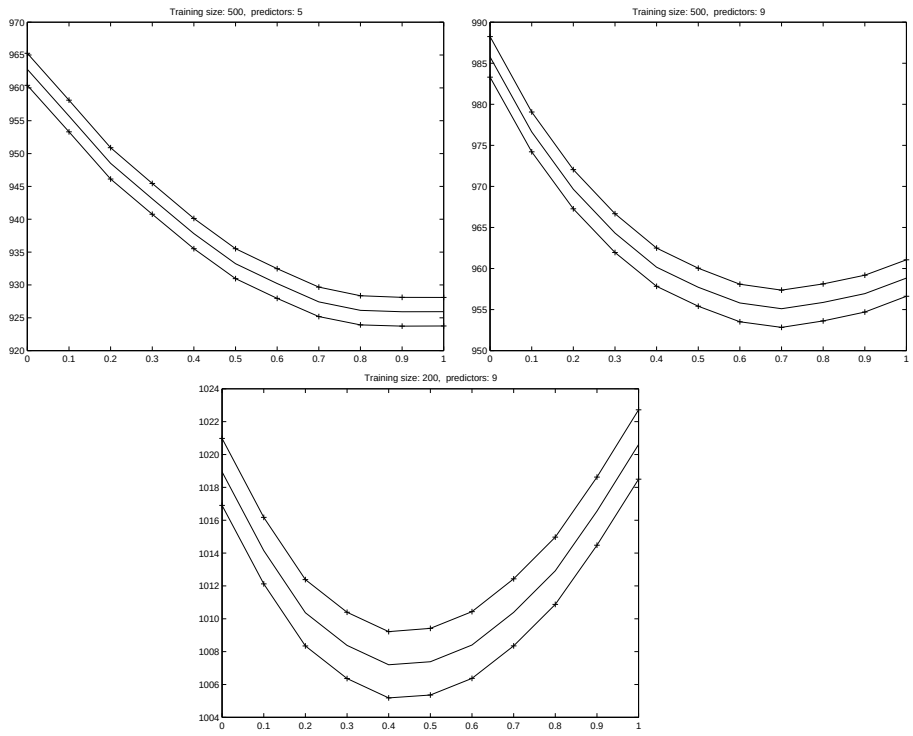

Figure 2: German Credit VW Misclassification as a function of power for various training set sizes and number of predictors

If we look at the actual gain from VWA on this dataset we see that while the results confirm our expectations, the difference between the future VW loss at p^* and the standard approaches (i.e. use $p = 0$ or $p = 1$) is no more than a few percent.

As we have mentioned in section 3.5, extreme distributions for the values cause both the VW variance and bias effects to be magnified. To illustrate this effect, we have run another analysis, but have changed the observation value for all 1000 observations from v_i to v_i^3.

The results of the model with 500 training size and nine predictors for this data can be seen in figure 3. As we can see, the optimal power is now around 0.4 (compared to $p^* \approx 0.7$ with values v_i) and the VW misclassification rate at both $p = 0$ and $p = 1$ is $12\% - 18\%$ worse.

Of course, our ability to evaluate the reduction in error from VWA is limited by our lack of knowledge of the best achievable VW error rate (the Bayes error rate). All reductions in VW error should theoretically be considered in

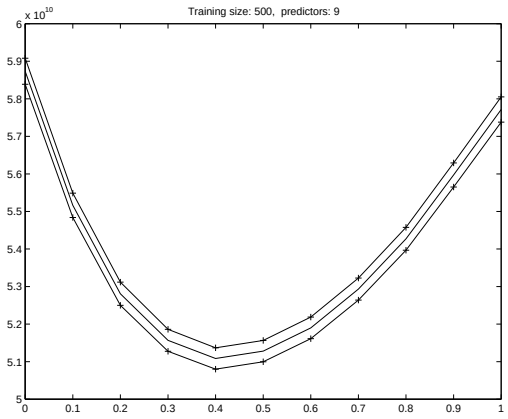

Figure 3: VW misclassification with training size 500, nine predictors and values v_i^3

terms of the remaining (reducible) error only. Unfortunately, this information is not available when dealing with real data, and so it is difficult to determine how "good" the reduction in VW error really is. We can only say that it is statistically significant and that the optimal power behaves as expected as the conditions and distribution of values change.

5.2 Census Data

This dataset is available from [Bay 99], called *ipums*:
http://kdd.ics.uci.edu/databases/ipums/ipums.html.

Similarly to the oft-used "adult" dataset, the task is to predict the income level ($\leq \$50K/year$ or $> \$50K/year$). Each observation has a sampling weight, indicating the actual size of population it represents. These weights can vary significantly - in this data they cover the range between 37 and 18000. This is a natural value indicator, since we obviously want to weigh our loss according to the sample weights, to represent "population" loss. We again consider minimizing VW future misclassification rate as the ultimate goal. We build logistic regression models with a range of values for p in training and examine their future VW misclassification.

Some examples of the future mean VW misclassification rate as a function of p for various numbers of predictors and training set sizes can be seen in figure 4.

As we can see, on this data $p = 0$ is clearly optimal for some models, while for others there does not seem to be much difference across the board for values of p. Notice also that the range within each graph does not exceed 3 units or 1% of the loss , while for different graphs the range varies significantly - up to 200 units! In other words, it seems that weighted analysis has almost no effect, and when it does it is to prefer the unweighted model. It seems that the situation

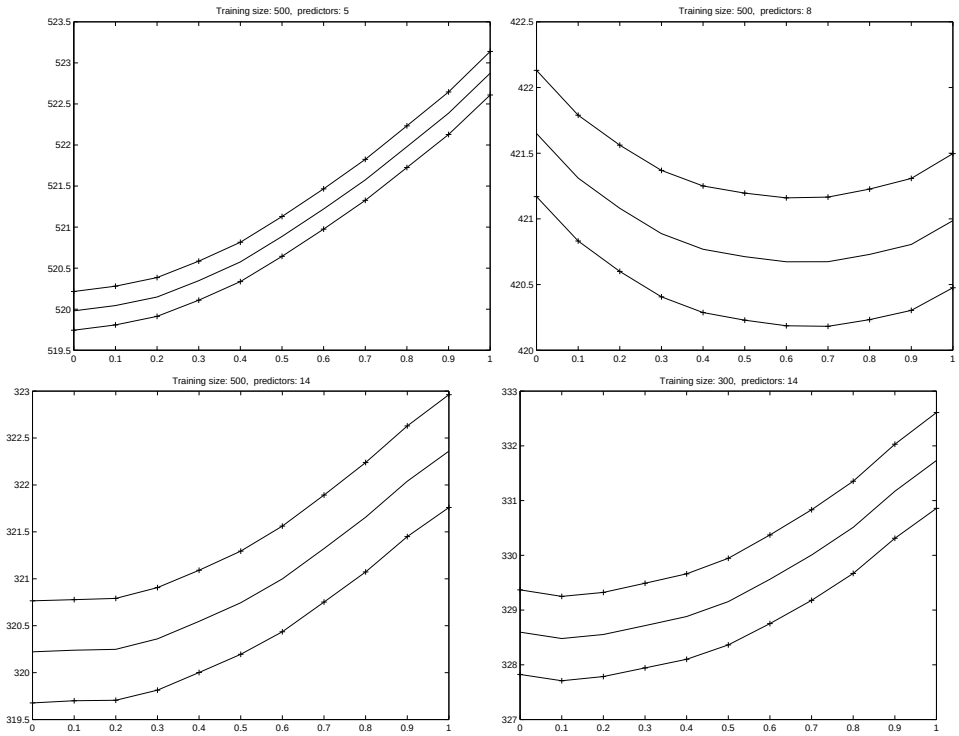

Figure 4: German Credit VW Misclassification as a function of power for various training set sizes and number of predictors

of proposition 3 is, to some extent, relevant.

6 Review of literature and discussion

[TO BE CONTINUED]

References

[Breiman 96] Breiman, L. (1996) *Bias, Variance, and Arcing Classifiers*, TR 460, Statistics Department, University of California.

[FHT 01] Friedman, J.H., Hastie, T. and Tibshirani, R. (2001) *Elements of Statistical Learning.* To appear.

[Friedman 97] Friedman, J.H. (1997) *On Bias, Variance 0/1 Loss and the Curse of Dimensionality*, In Knowledge Dis-

covery Journal Data Mining and Knowledge Discovery, 1:1, pp. 55-77, 1997.

[Fotherington 00] Fotheringham, A.S., Brunsdon, C., and Charlton, M.E. (2000) *Quantitative Geography*. London: Sage

[FED 00] Federal Trade Commission (2000).*Fair Information Practices in the Electronic Marketplace*. A report to Congress, Appendix A.

[IMPACT 00] IMPACT project (2000). *Behavioral Surveillance Surveys*, chapter 5.

[Potter 90] Potter, F.J. (1990) *A Study of Procedures to Identify and Trim Extreme Sampling Weights*. In Proc. 1990 Sect. Am. Stat. Assoc. on Survey Research Methods pp. 225-230.

[Korn and Graubard 95a] Korn, E.D., Graubard, B.I. (1995). *Analysis of Large Health Surveys: Accounting for Sample Design* JRSSA, 158:263-295.

[Korn and Graubard 95b] Korn, E.L., Graubard, B.I. (1995) *Examples of Differing Weighted and Unweighted Estimates from a Sample Survey*. The American Statistician, 49:291-295.

[Pfeffermann 93] Pfeffermann, D. (1993) *The Role of Sampling Weights when Modeling Survey Data*. International Statistical Review, 61(2):317-337.

[Pfeffermann et al. 98] Pfefferman, D., Skinner, C.J., Holms, D.J., Goldstein, H., Rasbash, J. (1998) *Weighting for Unequal Selection Probabilities*. JRSSB, 60:23-40.

[Bay 99] Bay, S. D. (1999). The UCI KDD Archive [http://kdd.ics.uci.edu]. Irvine, CA: University of California, Department of Information and Computer Science.

[Blake and Merz 98] Blake, C.L. and Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.

[Quinlan 93] Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

[BFOS 84] Breiman,L., Friedman, J.H., Olshen, R., Stone, C. (1984). *Classification and Regression Trees* ,Wadsworth,Belmont CA.

[McCullagh and Nelder 89] McCullagh, P., Nelder, J.A.(1989) *Generalized Linear Models*, 2nd Edition, Chapman & Hall.