“Coaching” variables for regression and classification

ROBERT TIBSHIRANI
Department of Preventive Medicine and Biostatistics
and Department of Statistics
University of Toronto
Toronto, Ontario
and
GEOFFREY HINTON
Department of Computer Science
University of Toronto

September 26, 1995

Abstract

In a regression or classification setting where we wish to predict Y from $x_1, x_2, \ldots x_p$, we suppose that an additional set of “coaching” variables $z_1, z_2, \ldots z_m$ are available in our training sample. These might be variables that are difficult to measure, and they will not be available when we predict Y from $x_1, x_2, \ldots x_p$ in the future. We consider two methods of making use of the coaching variables in order to improve the prediction of Y from $x_1, x_2, \ldots x_p$. The relative merits of these approaches are discussed and compared in a number of examples.

Keywords: regression, classification, missing data, mixtures of experts

1 Introduction

We consider the following problem: we have a response variable Y in a regression or classification setting that we wish to predict from a set of variables $\mathbf{x} = (x_1, x_2, \ldots x_p)$. In our training sample we have an additional set of variables $\mathbf{z} = (z_1, z_2, \ldots z_m)$. These variables are difficult to collect— for example an invasive medical test, or a financial indicator that is not obtainable quickly— and as a result they will not be available in the future for prediction of Y. Hence our present dataset consists of N measurements of $\mathbf{x}$, $\mathbf{z}$ and Y, but in the future we will have only $\mathbf{x}$ available for prediction of Y.

In this paper we address the following question: can we make use of $\mathbf{z}$ in our training sample, so as to improve future predictions of Y based on $\mathbf{x}$ alone? Breiman and Friedman (1994) have coined this the "coaching" problem, the idea being that $\mathbf{z}$ might be used to "coach" $\mathbf{x}$ in the art of predicting Y. We discuss two approaches to this problem, one called "mixture coaching" and the other due to Breiman and Friedman (1994) , which we call "response coaching".

A simple example shows that coaching is not always helpful. Suppose that $E(Y|x,z) = \alpha + \beta_1 x + \beta_2 z$, with $\text{var}(Y) = \sigma^2$. Then $E(Y|x) = \alpha + \beta_1 x + \beta_2 E(z|x)$. Now if z is independent of x, or even if $E(z|x)$ is a linear function of x, then $E(Y|x)$ is a linear function of x. It follows by the Gauss-Markov theorem that the least squares estimate of Y on x cannot be improved upon in a linear unbiassed sense by use of z.

More specialized models are needed to exploit coaching variables. In section 2 we describe two such models. Section 3 gives some examples with both real and simulated data. Our focus in this paper is on linear and generalized models: in section 4 we discuss some possibilities for nonparametric regression models.

2 Two coaching models

2.1 Mixture coaching

Let $f(B|A)$ denote the conditional density of B given A. Insight into the coaching problem can be gleaned from the relationship

$$f(Y|\mathbf{x}) = \int f(Y|\mathbf{x},\mathbf{z}) f(\mathbf{z}|\mathbf{x}) d\mathbf{z} \tag{1}$$

From (1) we see that it makes sense to make use of $\mathbf{z}$ if $f(Y|\mathbf{x},\mathbf{z})$ varies as a function of $\mathbf{z}$. That is, we can postulate a mixture model for Y given $\mathbf{z}$, the mixture components indexed by the value of the coach $\mathbf{z}$. To make use of $\mathbf{z}$ in this setting we also need some information in $\mathbf{x}$ about $\mathbf{z}$, that is $f(\mathbf{z}|\mathbf{x})$ must vary with $\mathbf{x}$.

Example 1. Two regression lines

Figure 1 shows a two regression line example. In both top panels $f(Y|x,z)$ is different for $z = 1$ and $z = 2$. In the top left panel, the distribution of $f(z|x) = .5$ independent of x so that it doesn't help to make use of z— with or without z we would estimate the mean of $f(y|x)$ by the regression line midway between the two lines.[1] The coach z tells us which regression line ("strategy")

[1] In a Bayesian sense, coaching is still helpful in this scenario. That is, the estimate which reports one regression line or the other with probability .5 is the Bayes estimate under the mixture model.

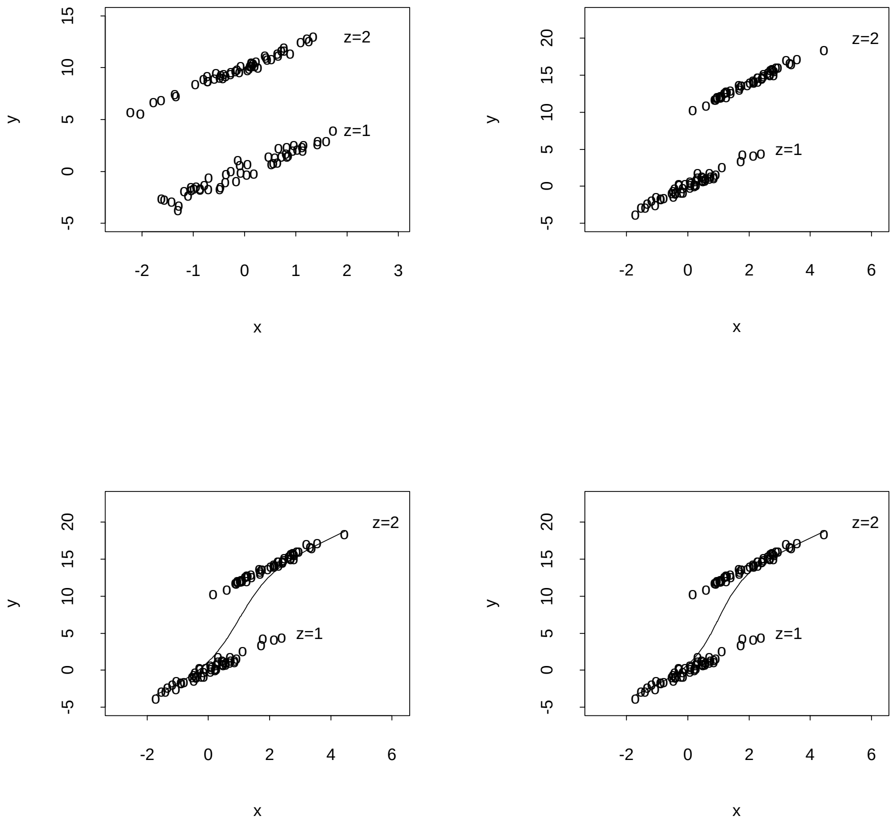

Figure 1: *Simulated data for mixture coaching. In the top left panel, coaching does not work since the distribution of z is independent of x. In the top right panel, coaching is possible: the bottom left panel shows the predicted values from mixture coaching using the true probabilities $f(z|x)$; in the bottom right the estimated values $\hat{f}(z|x)$ are used.*

to use, but there is no information in x to predict z in its absence. However in the top right panel $f(z|x)$ is not constant so that we can make use of z in coaching x. The bottom left panel shows the predictions when we use the actual probabilities $f(z|x)$; the bottom right shows the predictions when the probabilities $f(z|x)$ are estimated by the method given in this paper.

According to equation (1) we need to specify models for $f(Y|\mathbf{x},\mathbf{z})$ and $f(\mathbf{z}|\mathbf{x})$. The proposed model for $f(Y|\mathbf{x},\mathbf{z})$ is

$$Y = \mathbf{x}^T\beta(\mathbf{z}) + \epsilon \tag{2}$$

where ϵ has mean 0 and is independent of $\mathbf{x}$ and $\mathbf{z}$. That is, Y follows a linear regression in $\mathbf{x}$ for each value of $\mathbf{z}$, with the coefficients varying as a function of $\mathbf{z}$. A simple model would take $\beta(\mathbf{z}) = a + \gamma\mathbf{z}$, leading to a product interaction between $\mathbf{x}$ and $\mathbf{z}$. In a different context, Hastie and Tibshirani (1993) investigate more flexible models for $\beta(\mathbf{z})$, calling them "varying coefficient models".

For present purposes it is convenient to partition the $\mathbf{z}$ space into say K groups $A_1, A_2, \ldots K$. An attractive method for partitioning is recursive binary splitting, as used in the classification and regression tree methodology (CART; Breiman et al., 1984), and that is our approach here.

Recursive binary splitting requires a criterion for choosing a "best" split. In our problem we would like a split that gives a good fit for the piecewise linear model (2) and also one that is predictable from $\mathbf{x}$. Let $r(y,\mathbf{x},\boldsymbol{\beta}) = (y - \mathbf{x}\boldsymbol{\beta})^2$, and $d(y,p) = -y\log p - (1-y)\log(1-p)$. Suppose we have a node *parent* to be split into two nodes *son*1 and *son*2. Let $u_i = I(i \in son2)$. Define

$$\begin{aligned} rss &= \sum_{i\in son1} r(y_i,\mathbf{x}_i,\hat{\boldsymbol{\beta}}_1) + \sum_{i\in son2} r(y_i,\mathbf{x}_i,\hat{\boldsymbol{\beta}}_2) \\ dev &= \sum_{i\in parent} d(u_i,\hat{p}_i) \end{aligned} \tag{3}$$

Here $\hat{\beta}$ is the least squares estimate of y on $\mathbf{x}$ in the *parent* node (and similarly for $\hat{\beta}_1$ and $\hat{\beta}_2$), and

$$\hat{p}_i = \hat{p}(x_i) = 1/[1 + \exp(-\hat{\theta}^T\mathbf{x}_i)] \tag{4}$$

where $\hat{\theta}$ maximizes $\sum_{i\in parent} d(u_i,p_i)$. Then we choose the split to minimize

$$cost = \Delta rss + \lambda \cdot dev \tag{5}$$

where $\lambda > 0$ is a parameter that balances the two objectives. The best value for λ may be derived from a cross-validation estimate of the error in predicting Y from $\mathbf{x}$.

The terminal nodes of the tree give a partition $A_1, A_2, \ldots A_K$ of the training data based in values of $\mathbf{z}$. In our implementation, for simplicity we limit the number of terminal nodes in the tree to some small fixed number, say 2 or 3, and do not use bottom-up pruning.

Notice that in this process we have in effect estimated two models from the training data: a model for predicting Y from $\mathbf{x}$ in different partitions A_j of the $\mathbf{z}$ space, and a model for predicting the partition membership from $\mathbf{x}$. We call the first model the *coaching model* for Y, and the second one the *strategy model* for $\mathbf{z}$. The coaching model tells us how to predict Y from $\mathbf{x}$ for each of the "strategies" A_j while the strategy model predicts the strategy to use when $\mathbf{z}$ (and hence the partition membership) is unknown.

In our particular construction, the coaching and strategy models are based on the same binary splitting of the training data. This is natural when recursive binary splitting is used to form the partition. However this need not be true if other approaches are used to estimate the partition.

To predict Y from a value $\mathbf{x}_0$, we estimate the probabilities $\pi_k(\mathbf{x}_0) = f(A_k|\mathbf{x}_0)$ from the strategy model: these are just products of the conditional probabilities $\hat{p}(x)$ at the splits defining each terminal region. Then we compute the discrete analog of (1):

$$\hat{f}(Y|\mathbf{x}_0) = \sum_{k=1}^{K} \hat{\pi}_k(\mathbf{x}_0)\hat{f}(Y|\mathbf{x}_0, A_k)$$

Specifically, for the linear regression model we obtain $\hat{E}(Y|\mathbf{x}_0) = \sum_{k=1}^{K} \hat{\pi}_k(\mathbf{x}_0)\mathbf{x}_0^T\hat{\beta}_k$.

Remark A. To extend the mixture coaching model to generalized regression models, we take the linear part of the model to be $\eta = \mathbf{x}^T\beta(\mathbf{z})$. Consequently we fit a generalized regression model to the data in each of the nodes of the strategy tree.

Remark B. In practice we have found that the procedure is quite insensitive to the value chosen for λ. A value of 0 is often as good as values > 0 and in fact, zero is the value used in most of our examples. It is possible, however, to construct examples in which a value $\lambda > 0$ is optimal (Example 2 below).

Remark C. The mixture coaching technique has some similarity to the "surrogate variable" facility in the CART procedure. Here we are splitting on the $\mathbf{z}$ variables, and seeking surrogates among the $\mathbf{x}$ variables. However rather than surrogate variables, we actually obtain probabilities of going left or right at each split. And further, we choose the split based on the fit *and* predictability of the split from $\mathbf{x}$.

Remark D. It might be that some of the coaching variables are missing in some of the training cases, for example if a patient refused an invasive medical test. This would create no difficulty for the mixture coaching procedure, as we simply use the cases with complete data at each node. This is analogous to surrogate variable approach used in CART. Similarly, if some but not all of the $\mathbf{z}$ variables were available for prediction, we would simply use their observed vales in the strategy tree.

Remark E. Using a coaching variable in this way is analogous to fitting a mixture density with labelled data. Hence a rough idea of the possible gains in efficiency over the use of unlabeled data can be obtained from results in density estimation, for example Hosmer and Dick (1974). Their work also suggests that substantial gains can be achieved even from partially labelled data. Such data can be handled in the coaching problem by the method described in Remark D.

2.2 A network representation of mixture coaching

The mixture coaching model is closely related to the "mixtures of local experts" approach of Nowlan (1991), Jacobs *et. al.* (1991) and Jordan and Jacobs (1994). The main distinction is that in the coaching problem, the label of the strategy or expert is known in the training set while in the mixture of experts method, the label of the expert in the training set is unknown. and is estimated as a linear function of $\mathbf{x}$ from the data.

Following up on this connection, Figures 2 and 3 show a network representation of the mixture coaching procedure. Details are given in the figure captions.

2.3 Response coaching

Another form of equation (1) is

$$f(Y|\mathbf{x}) = \int f(Y, \mathbf{z}|\mathbf{x}) d\mathbf{z}. \tag{6}$$

This suggests that we can use $\mathbf{Z}$ as a coach by jointly predicting both Y and $\mathbf{Z}$ from $\mathbf{x}$. Suppose that $f(Y, \mathbf{Z}|\mathbf{x})$ is indexed by parameters $\theta = (\theta_1, \theta_2)$ and $\boldsymbol{\gamma}$. The simplest models would specify independence of Y and $\mathbf{Z}$ given $\mathbf{x}$:

$$f_{\theta,\boldsymbol{\gamma}}(Y, \mathbf{Z}|\mathbf{x}) = f_{\theta_1,\boldsymbol{\gamma}}(Y|\mathbf{x}) \cdot f_{\theta_2,\boldsymbol{\gamma}}(\mathbf{Z}|\mathbf{x}) \tag{7}$$

Given estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\boldsymbol{\gamma}}$, equation (6) becomes

$$\hat{f}(Y|\mathbf{x}) = f_{\hat{\theta}_1,\hat{\boldsymbol{\gamma}}}(Y|\mathbf{x}) \tag{8}$$

and we would predict Y from $\mathbf{x}$ using $f_{\hat{\theta}_1,\hat{\boldsymbol{\gamma}}}(Y|\mathbf{x})$. We call (7) and (8) the *response coaching* model. In a sense, the response coaching model is a special case of the mixture coaching procedure. There is only one strategy, and we do not partition the $\mathbf{z}$ space but treat $\mathbf{z}$ as is. Finally, we use a common set of parameters in the coaching and the strategy models.

It is the presence of a common set of parameters $\boldsymbol{\gamma}$ in the models for Y and $\mathbf{Z}$ that leads to any potential benefit from $\mathbf{Z}$. For example if we specified separate linear regression models for Y and $\mathbf{Z}$, the prediction of Y would be unchanged by inclusion of $\mathbf{Z}$. However consider a model of the form

$$Y = \sum_{j=1}^{J} \theta_{1j} h(\mathbf{x}, \boldsymbol{\gamma}_j) + \epsilon_1$$

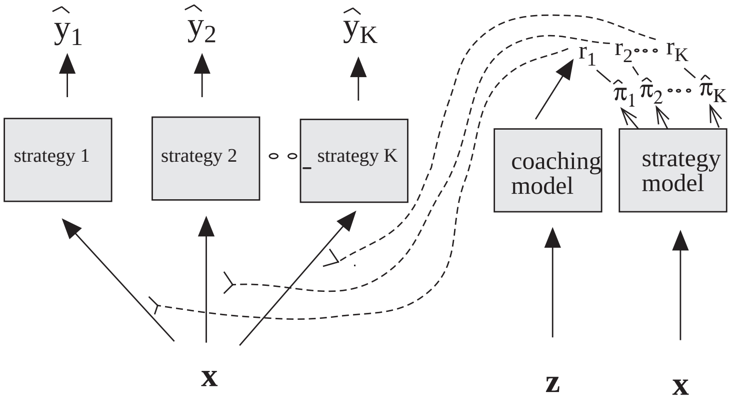

Figure 2: Fitting of the mixture coaching model. The coaching model uses $\mathbf{z}$ to partition the data into K regions. The binary variables $r_1, r_2, \ldots r_K$ are indicator variables for the K regions. In each of the regions we use a different strategy (or "expert"), specifically a linear model in $\mathbf{x}$. The strategy model estimates probabilities of partition membership π_k, based on $\mathbf{x}$.

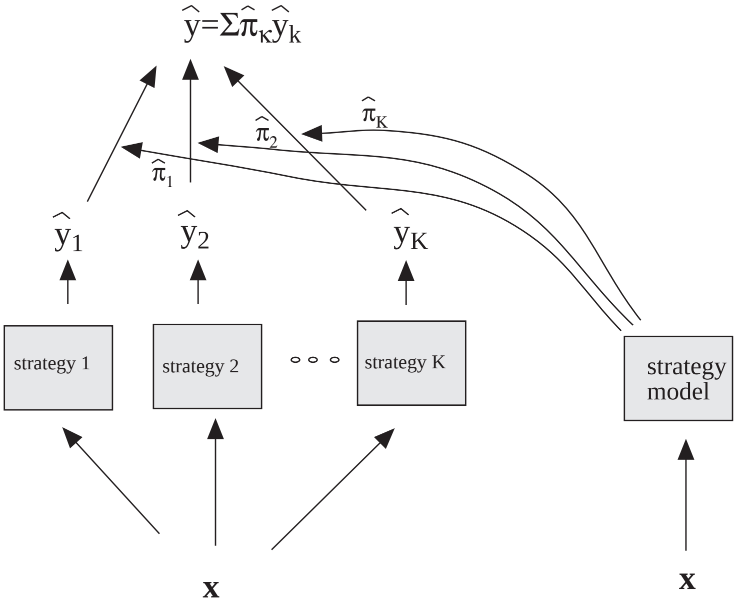

Figure 3: Prediction from the mixture coaching model. The strategy model estimates the probabilities π_k with which to combine the outputs of each of the strategies, into the fitted value $\hat{y}$. Alternatively, we can view the π_k as mixing proportions in a mixture of Gaussians model for Y given $\mathbf{x}$.

$$\mathbf{Z} \;=\; \sum_{j=1}^{J} \theta_{2j} h(\mathbf{x}, \boldsymbol{\gamma}_j) + \epsilon_2 \tag{9}$$

where ϵ_1 and ϵ_2 are independent. Here Y and $\mathbf{Z}$ share the basis functions $h(\mathbf{x}, \boldsymbol{\gamma}_j)$ and this leads to potential improvements due to the presence of $\mathbf{Z}$.

Consider the simple case where $\mathbf{Z}$ is a scalar, $J = 1$, and $h(\mathbf{x}, \boldsymbol{\gamma}) = \boldsymbol{\gamma}^T\mathbf{x}$. For identifiability assume $||\boldsymbol{\theta}|| = 1$ where $\boldsymbol{\theta} = (\theta_1, \theta_2)$. Assume $\mathrm{var}(\epsilon_j) = \sigma_j^2$. Then the maximum likelihood estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ can be derived from a canonical correlation analysis of y and $\mathbf{x}$. Here is a convenient form of the solutions. Let $P_{\mathbf{X}}$ denote the projection matrix onto the columns of the observed $\mathbf{x}$ values and let $\mathbf{W}$ be the $n \times 2$ matrix of observed y and z values. Then $\hat{\boldsymbol{\theta}}$ is the leading eigenvector of $\mathbf{W}^T P_{\mathbf{X}} \mathbf{W}$, and $\hat{\boldsymbol{\gamma}}$ is the coefficient of the least squares regression of $\hat{\boldsymbol{\theta}}^T \mathbf{W}$ on $\mathbf{x}$.

The variance of $\hat{\boldsymbol{\gamma}}$ is complicated since $\hat{\boldsymbol{\theta}}$ is a nonlinear function of $\mathbf{x}$. However if we assume that $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ the true value, it is easy to show that $\hat{\boldsymbol{\gamma}} = \theta_1 \hat{\boldsymbol{\gamma}}^y + \theta_2 \hat{\boldsymbol{\gamma}}^z$ where $\hat{\boldsymbol{\gamma}}^y$ is the least squares estimate of y on $\mathbf{x}$ and similarly for $\hat{\boldsymbol{\gamma}}^z$. Suppose $\sigma_2^2 = c\sigma_1^2$. It follows that

$$\mathrm{var}(\hat{\boldsymbol{\gamma}}) = (\theta_1^2 + c\theta_2^2)\mathrm{var}(\hat{\boldsymbol{\gamma}}^y) \tag{10}$$

If for example $\theta_1 = \theta_2 = 1/\sqrt{2}$ and $c = 1/4$, then $(\theta_1^2 + c\theta_2^2) = .53$. Hence the use of the coach z would reduce the variance of $\hat{\boldsymbol{\gamma}}$ by approximately half. The simulation results in Example 3 below suggest that the actual gains are somewhat less than this.

Remark F. Another instance of the response coaching model would be parallel regression trees for Y and Z, each tree being a function of $\mathbf{x}$. The trees would have the same splits. i.e. basis functions $h(\mathbf{x}, \boldsymbol{\gamma}_j)$, but different estimated values in each node (parameters θ_{1j} and θ_{2j}). In this way, z could coach $\mathbf{x}$ in making useful partitions of the $\mathbf{x}$ space for predicting Y.

3 Examples

Example 1 continued.

Example 1 is designed for the mixture coaching procedure. For the data in the top right panel of Figure 1, the mixture coaching tree simply splits z into the two groups $z = 1$ and $z = 2$. Figure 1 shows the test sample results from this model. There were 200 training observations and 500 test observations in each of 20 simulations. The results in the first boxplot used the actual value $f(z|x)$ from the model, while in the second one we estimated $f(z|x)$ from the strategy model. Both outperform the "no coach" model (simple least squares regression). The (linear) response coaching model gives identical results to the no coaching model since the rank of the regression matrix is only one.

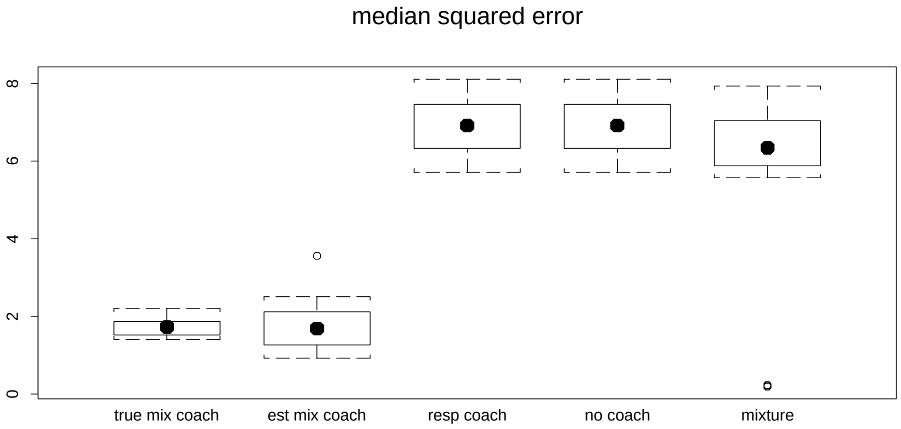

Figure 4: *Results for Example 1 (two regression lines)*

The rightmost boxplot labelled "mixture" refers to the fitting of a mixture of 2 regression lines, without any coaching. We started with a random allocation of the points to one of two groups and then used a standard EM algorithm to estimate the intercepts and slopes of the lines, and the mixing proportion for each observation. Five different random seeds were used, and the one giving highest log-likelihood was selected. To predict $\hat{y}$ for a gvien value x_0, we found the nearest neighbour x in the training set, and used the estimates for x. The results show that uncoached fitting of mixtures does about as well as simple least squares regression. The problem is that for very few starting values does the algorithm converge to the desired two regression line solution. For almost all starting values both regression lines converge to lines that are very close to the overall regression line. Of course one could take many more starting values, but this difficulty would be exacerbated with high dimensional $\mathbf{x}$ and many mixture components.

Example 2: Multiple coaches
In this example we generated 50 observations from the model

$$\begin{aligned} z_j &\sim Unif(0,2),\ j=1,2,3 \\ x &= z_1+\epsilon_1 \\ \alpha &= \beta = z_1+z_2+z_3 \\ y &= \alpha+\beta x+\epsilon_2 \end{aligned} \tag{11}$$

where the components of ϵ_1 and ϵ_2 are independent normal $N(0,.25)$. All three z_j's affect the mean of y, but only z_1 is predictable from x. Figure 5 shows

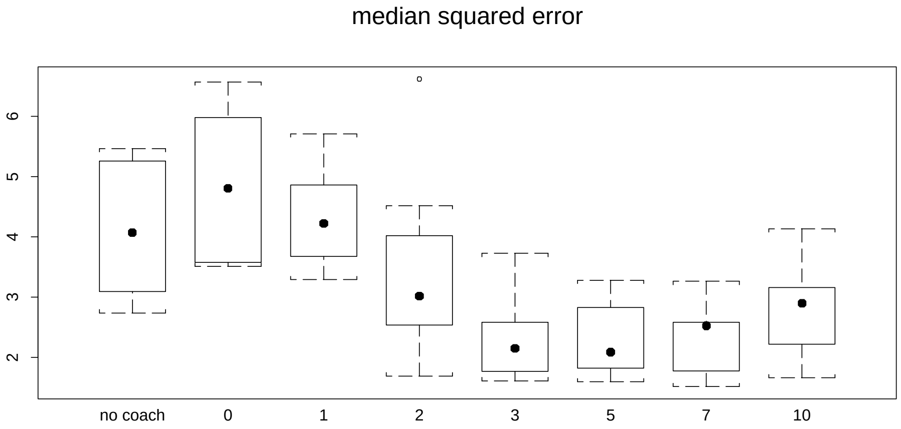

Figure 5: *Results for Example 2 (multiple coaches). Median squared error for no coach model, and mixture coaching with* $\lambda = 0, 1, \ldots 10$.

the test error (over 10 simulations) from the mixture coaching procedure as a function of the cost parameter λ. For values of $\lambda > 0$, the criterion takes account of the predictability of the coach and produces a small improvement. In this example we limited the number of terminal nodes in the coaching tree to 2, i.e. one split. When more splits were allowed, the effect of non-zero values of λ disappeared.

Example 3: Bivariate reduced rank model

In this example we generated data from the model

$$\begin{aligned} y &= 2\mathbf{a}^T\mathbf{x} + \epsilon_1 \\ z &= 2\mathbf{a}^T\mathbf{x} + 0.25 \cdot \epsilon_2 \end{aligned} \tag{12}$$

where the components of $\mathbf{x}, \epsilon_1$ and ϵ_2 are independent standard normal and $\mathbf{a} = (1, -2, 1)$. This model might reflect a situation where z is a "gold standard" measurement and y is a noisy version of z. The results of 20 simulations from this model are shown in Figure 6.

This example is of course perfectly suited to the response coach model. However Figure 6 shows that the gains over the no coach (simple regression) approach are only modest.

Example 3: NO_x data

Cleveland *et al.* (1991) examine 88 observations on the exhaust from an engine fueled by ethanol. The response variable, denoted by NO_x, is the concentration of nitric oxide and nitrogen dioxide, normalized by the work load of

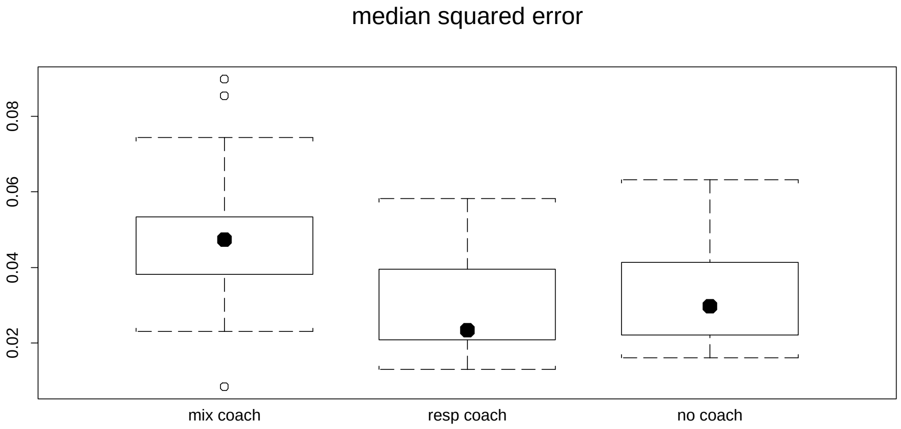

Figure 6: *Results for Example 3 (bivariate normal model)*

the engine. The two predictors are equivalence ratio E, a measure of the fuel/air mixture, and the compression ratio C of the engine.

Figure 7 gives a plot of the data. The broken lines show the fitted linear regressions of NO_x on C in 4 nonoverlapping ranges of E. Within each range of E, a linear model in C seems to fit well. But as E varies, both the intercept and slope of the line vary. This suggests that E might act as an effective mixture coach for C. We tried divided the data randomly into training and test sets of size 44, and applied the coaching model of the previous section. This was done for 5 random divisions, and a summary of the results is shown in Figure 8.

Mixture coaching by E clearly helps in the prediction of NO_x from C. The (linear) response coaching model gives identical results to the no coaching model since the rank of the regression matrix is only one.

Example 4: Detection of Muscular Dystrophy Carriers

These data are taken from Andrews and Herzberg (1985). They consist of the enzyme measurements on 209 female relatives of boys with Duchenne Muscular Dystrophy (DMD). The females are either DMD carriers or normals. The overall objective is to predict DMD from the enzyme levels. There are four enzyme measurements - CK and H, which are inexpensive to collect, and PK and LD, which are more costly. According to Andrews and Herzberg, one of the important questions was whether the second two enzymes increase the detection rate in an important way.

The women's age was also available, and we included it in the analysis. After deleting incomplete observations there were 194 cases– 127 normals and 67 carriers. Some women were measured more than once on different days and

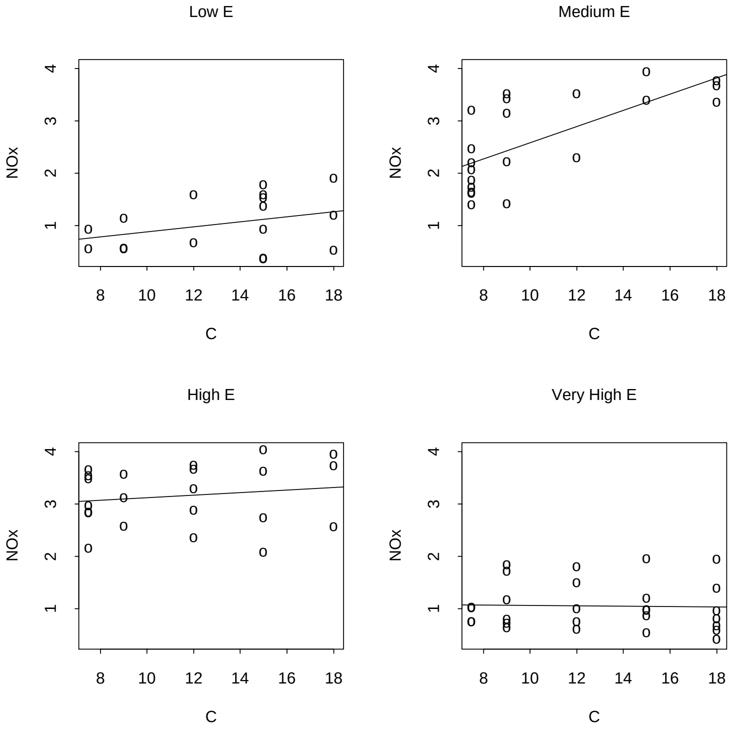

Figure 7: *Plots of NO_x versus C for low, medium, high and very high values of E. Included in each panel are least squares regression lines*

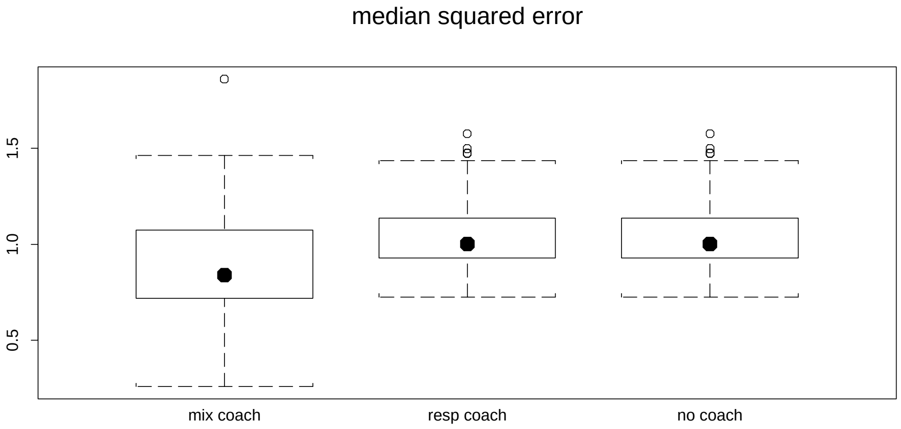

Figure 8: *Results for Example 3 (NO_x data)*

Table 1: *Linear logistic analysis of DMD data*

Variable	Estimate	Standard Error	Z score
age	0.16	0.05	3.53
CK	2.88	0.75	3.82
H	0.09	0.03	3.17
PK	1.88	0.80	2.34
LD	0.01	0.01	1.92

we retained the replicates. Table 1 shows the results of a linear logistic analysis of these data.

The deviance with and without (PK, LD) was 81.8 and 94.1, suggesting that PK and LD significantly improve the detection rate. To investigate this , we randomly divided the data into two halves, trained on one half and predicted the other half. The results of 20 such random divisions of the data are shown in Table 2, including linear logistic models (lines 1 and 2) a linear logistic model coached by PK and LD (line 3), and a linear response coaching model (line 4). We see that the PK and LD enzymes do not improve upon age, CK and H. Given this fact, it is not surprising that PK and LD are not effective coaches for CK and H.

Table 2: *Results of 20 validation draws for the DMD data*

Model	Median deviance (se)	Error rate(se)
(1) All variables	0.51 (.02)	0.09 (.01)
(2) Age, CK, H	0.50 (.03)	0.10 (.01)
(3) Mixture coaching by PK, LD	0.54 (.06)	0.10 (.01)
(4) Response coaching by PK, LD	0.56 (.02)	0.10 (.01)

4 Discussion

In this paper we have investigated the effectiveness of two different approaches to the utilization of "coaching" variables. The two approaches- mixture and response coaching, seem to complement each other and in some cases can improve prediction accuracy.

Our focus has been on linear and generalized models: however extensions to more flexible nonparametric regression models are possible. For example, in the mixture coaching procedure one could allow splits on $\mathbf{x}$ as well as $\mathbf{z}$, which would make the procedure more like the CART method. In the response coaching model, one could use adaptively selected basis functions in the models for Y and $\mathbf{Z}$: this approach is explored in Breiman and Friedman (1994).

It would be interesting to explore applications of coaching methods to financial forecasting and other time series problems.

Acknowledgments

We would like to thank Brad Efron, Andrey Feuerverger, Trevor Hastie, Michael Jordan, Jerry Lawless, Michael LeBlanc and Jim Zidek for helpful discussions, and Jerome Friedman for introducing us to the coaching problem. Both authors were supported by the Natural Sciences and Engineering Research Council of Canada.

References

Andrews, D. & Herzberg, A. (1985), *Data*, Springer-Verlag, Berlin.

Breiman, L. & Friedman, J. (1994), Multivariate multiple shrinkage, Poster at Snowbird Neural Nets Conference.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth.

Cleveland, W., Grosse, E., Shyu, W. & Terpenning, I. (1991), Local regression models, *in* J. Chambers & T. Hastie, eds, 'Statistical models in S', Wadsworth.

Hastie, T. & Tibshirani, R. (1993), 'Varying coefficient models (with discussion)', *J. Royal. Statist. Soc. B* **55**, 757–796.

Hosmer, D. & Dick, N. (1974), 'Information and mixtures of two normal distributions', *J. Statist. Comput. Simul.* pp. 995–1006.

Jacobs, R., Jordan, M., Nowlan, S. & Hinton, G. (1991), 'Adaptive mixtures of local experts', *Neural computation* **3**, 79–87.

Jordan, M. & Jacobs, R. (1994), 'Hierachical mixtures of experts and the em algorithm', *Neural computation* **6**, 181–214.

Nowlan, S. (1991), Soft competition and adaptation, Technical report, PhD. thesis, Comp. Sci., Carnegie Mellon University.