A Unified Bias-Variance Decomposition and its Applications

Pedro Domingos
Department of Computer Science and Engineering
University of Washington

ICML 2000

Presented by Charles Rosenberg
November 13, 2000

Presentation Outline

- Introduction
- Definitions
- Bias-Variance Decomposition for Zero-One Loss
- Results
- Summary

Bias and Variance

Problem Setup

Given a trained model from a machine learning algorithm, the goal is to better understand what factors (including the data set specifics and the algorithm's parameters) affect the algorithm's performance and why.

Bias-Variance Decomposition

To understand measured loss, it is often decomposed into three terms:

$$Expected\ Loss = f(Bias, Variance, Noise)$$

Zero-One Loss

What about classification problems? What is the right way to handle loss defined as 1 when a prediction is correct and 0 when a prediction is not?

Basic Definitions

$\{(x_1, t_1), ..., (x_n, t_n)\}$ is a training set drawn from D, a set of training sets of size n

$y = f(x)$ is the prediction generated by our trained model f

t is the true target value for a specific test example x, typically determined by a non-deterministic function of the value of x

$L(t, y)$ is the ***loss function***, for example:

squared loss: $L(t, y) = (t - y)^2$

absolute loss: $L(t, y) = |t - y|$

zero-one loss: if $y = t$ then $L(t, y) = 0$, else $L(t, y) = 1$

$E_{D,t}[L(t, y)]$ is the expected loss over all target values and training sets

Prediction Type Definitions

$y_* = \arg\min_{y_*} E_t\left[L(t, y_*)\right]$

- It is defined to be the ***optimal prediction*** for a specific example x.
- It is the lowest loss prediction for a specific value of x irrespective of our model.

$y_m = \arg\min_{y'} E_D\left[L(y, y')\right]$

- It is defined to be the ***main prediction*** for a specific value x, a specific loss function L and a set of training sets D.
- It is a value that differs least from all other predictions Y according to L.
- It is not necessarily a member of Y.

Key Definitions

$B(x) = L(y_*, y_m)$ is defined to be the ***bias of a learner***. It is the loss incurred by the main prediction relative to the optimal prediction.

$V(x) = E_D\left[L(y_m, y)\right]$ is defined to be the ***variance of a learner***. It is the average loss incurred by the models' predictions relative to the optimal prediction.

$N(x) = E_t\left[L(t, y_*)\right]$ is defined to be the ***noise of a learner***. It is the unavoidable component of loss, incurred independently of the learning algorithm.

If the loss function is non-negative then bias and variance are also non-negative.

Proposed Bias-Variance Decomposition

$E_{D,t}\left[L(t,y)\right]$ is the expected loss for a specific example x with true prediction t given that a learner predicts y given a training set from D.

The decomposition is as follows:

$$E_{D,t}\left[L(t,y)\right] = c_1 E_t\left[L(t,y_*)\right] + L(y_*,y_m) + c_2 E_D\left[L(y_m,y)\right]$$
$$E_{D,t}\left[L(t,y)\right] = c_1 N(x) + B(x) + c_2 V(x)$$
$$\Big(E_{D,t,x}\left[L(t,y)\right] = E_x\left[c_1 N(x)\right] + E_x\left[B(x)\right] + E_x\left[c_2 V(x)\right]\Big)$$

For squared loss, standard definition is recovered with $c_1 = c_2 = 1$, $y_* = E_t[t]$ and $y_m = E_D[y]$:

$$E_{D,t}\left[(t-y)^2\right] = E_t\left[(t-E_t[t])^2\right] + (E_t[t]-E_D[y])^2 + E_D\left[(E_D[y]-y)^2\right]$$

Decomposition for Two Class Zero-One Loss

$$E_{D,t}[L(t,y)] = c_1 E_t[L(t,y_*)] + L(y_*,y_m) + c_2 E_D[L(y_m,y)]$$

$$E_{D,t}[L(t,y)] = c_1 N(x) + B(x) + c_2 V(x)$$

Let $P_D(y = y_*)$ be the probability that the learner predicts the optimal class for x over the training sets in D.

$$c_1 = 2P_D(y = y_*) - 1$$

if $y = y_*$ $(B(x) = 0)$ $c_2 = 1$, otherwise $(B(x) = 1)$ $c_2 = -1$

$$E_{D,t}[L(t,y)] = c_1 N(x) + B(x) + (1 - B(x))V_u(x) - B(x)V_b(x)$$

Margins and Bias-Variance

Two class margin definition:

$$M(x) = P_D(y = t) - P_D(y \neq t)$$

Margin defined in terms of zero-one bias and variance:

$$M(x) = c \times [2B(x) - 1][2V(x) = 1]$$

Where $c = +1$ if $y_* = t$ and $c = -1$ otherwise.

Similarly:

$$B(x) = \frac{1}{2}[1 + c \times sign(M(x))]$$
$$V(x) = \frac{1}{2}[1 + c \times |M(x)|]$$

Experimental Set Up

- 30 Data Sets from the UCI Repository.
- Assumed that $N(x) = 0$.
- Data divided into a training set $\left(\frac{2}{3}\right)$ and test set $\left(\frac{1}{3}\right)$.
- Generate 100 bootstrap samples from each training set.

Decision Tree Results

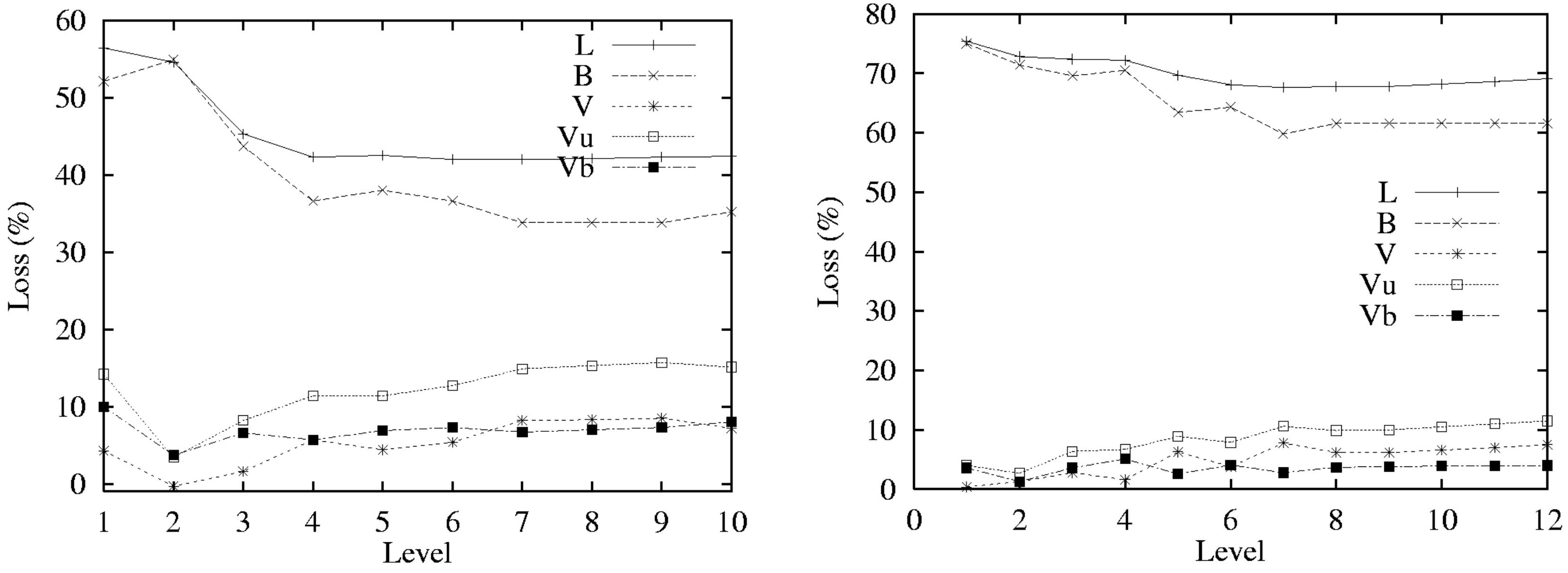

Figure 1. Effect of varying the number of levels in C4.5 trees: glass and primary tumor.

Boosted Decision Tree Results

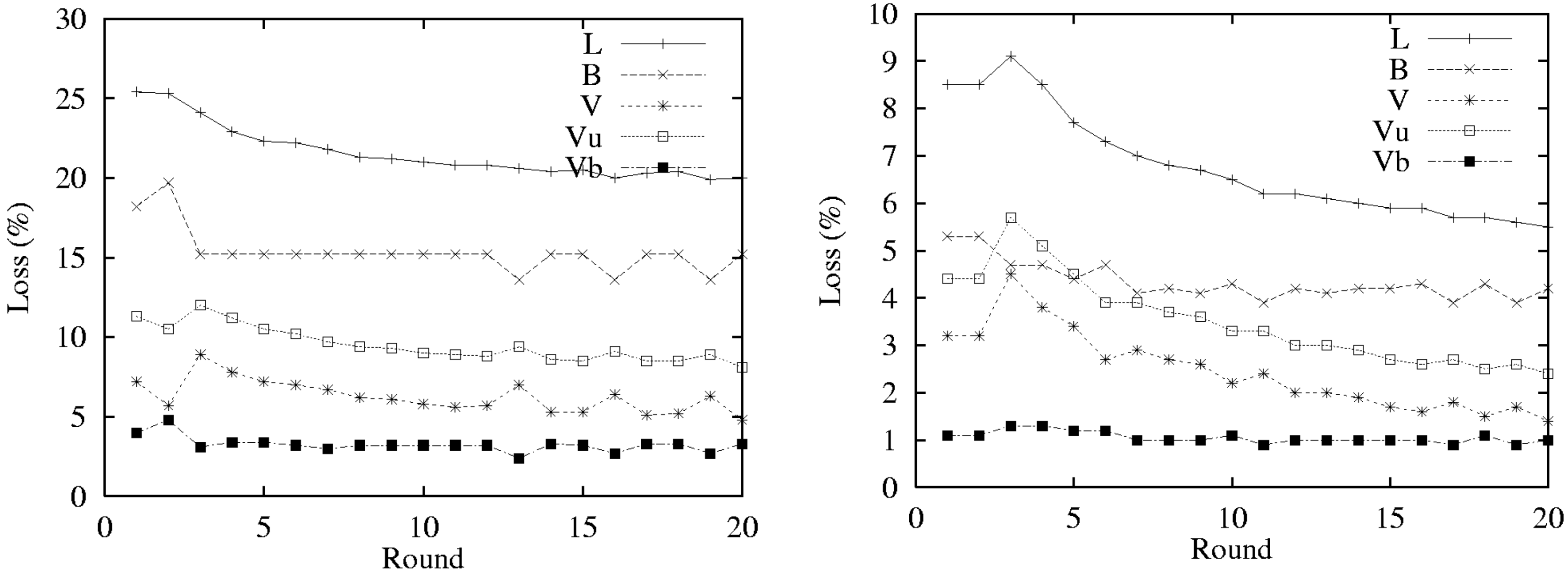

Figure 2. Effect of boosting on C4.5: audiology and splice junctions.

K-Nearest Neighbor Results

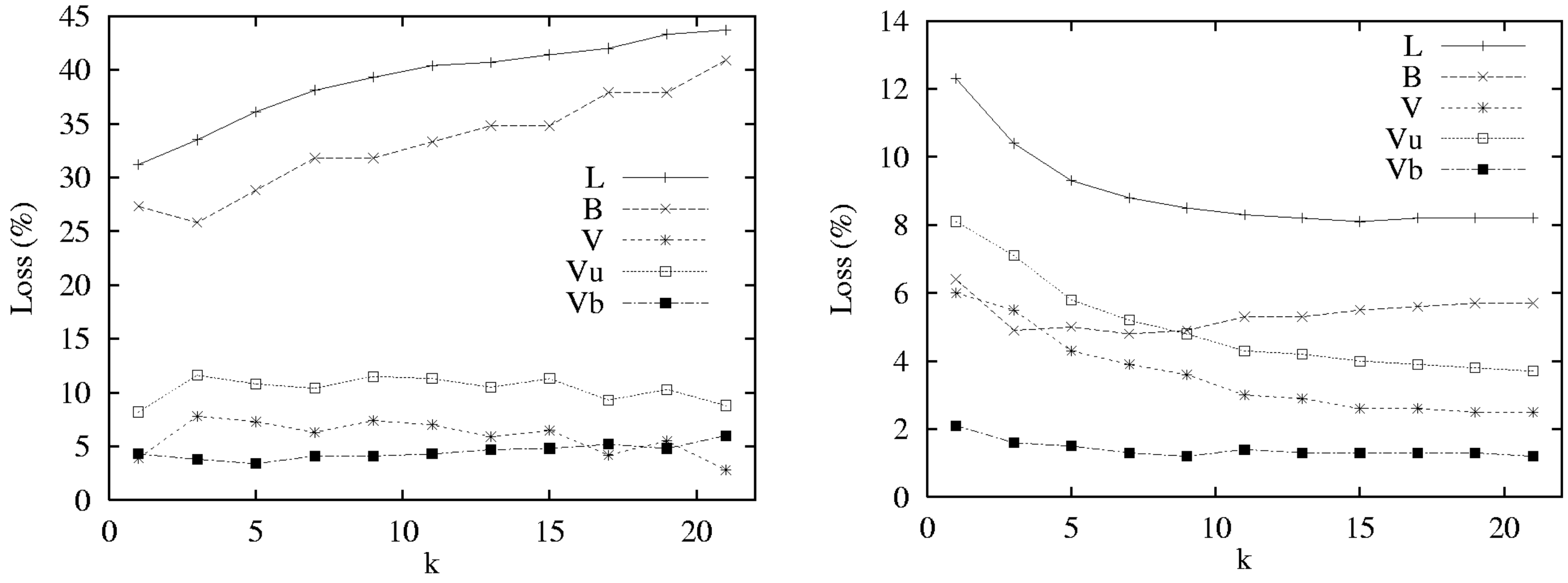

Figure 3. Effect of varying k in k-nearest neighbor: audiology and chess.

Summary

- Proposes a unified definition of bias and variance which is applicable to any loss function.
- Proposed decomposition specializes to standard for squared loss and avoids previous difficulties for zero-one loss.
- Slightly unusual in that components of the decomposition are not always additive.
- The main limitation is that certain loss functions (e.g. absolute loss) can not be decomposed according to Equation 1.
- Code is available from his web site.