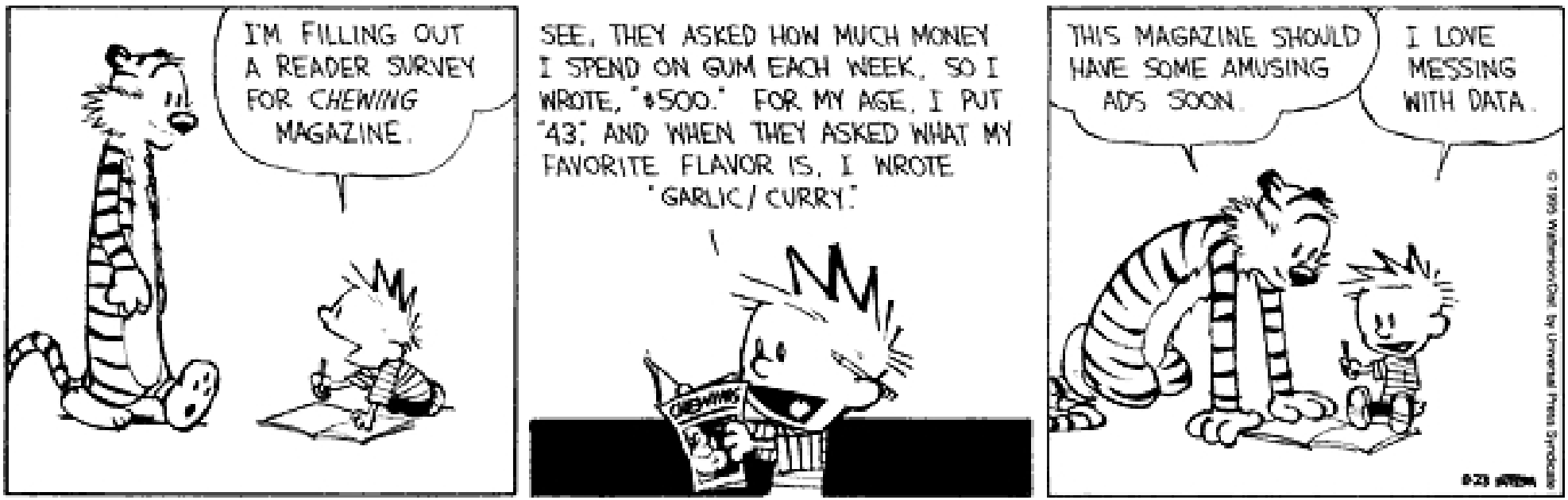
STAT406- Methods of Statistical Learning Lecture 2

Matias Salibian-Barrera

UBC - Sep / Dec 2018


I'M FILLING OUT A READER SURVEY FOR CHEWING MAGAZINE.
SEE, THEY ASKED HOW MUCH MONEY I SPEND ON GUM EACH WEEK, SO I WROTE, "$500." FOR MY AGE, I PUT "43." AND WHEN THEY ASKED WHAT MY FAVORITE FLAVOR IS, I WROTE "GARLIC/CURRY."
CHEWING
THIS MAGAZINE SHOULD HAVE SOME AMUSING ADS SOON.
I LOVE MESSING WITH DATA.
© 1989 Watterson

Predictions

Mean squared prediction error (MSPE)

$$Y \longleftrightarrow \hat{f}_n(\mathbf{X})$$

$$\left(Y - \hat{f}_n(\mathbf{X})\right)^2 \quad ?$$

$$E\left[\left(Y - \hat{f}_n(\mathbf{X})\right)^2\right] \quad ?$$

What are we “averaging” over?
What is random?

Predictions

- What we want, typically, is

$$E_{(Y^*,\mathbf{X}^*)}\left[\left(Y^* - \hat{f}_n\left(\mathbf{X}^*\right)\right)^2\right]$$

where $(Y^*, \mathbf{X}^*)$ are new, future observations, not used when computing ("training") $\hat{f}_n$.

Predictions

If we assume that $Y = f(X) + \epsilon$, then

$$E_{(Y^*,\mathbf{x}^*)}\left[\left(Y^* - \hat{f}_n(\mathbf{X}^*)\right)^2\right] =$$

$$E_{(Y^*,\mathbf{x}^*)}\left[\left(f(\mathbf{X}^*) - \hat{f}_n(\mathbf{X}^*)\right)^2\right] + \mathsf{V}(\epsilon)$$

- what assumptions are needed for this to be true?
- is it still true if I look at predictions for a single & fixed $\mathbf{X}_0$?

Predictions

- What we want

$$E_{(Y^*,\mathbf{x}^*)}\left[\left(Y^* - \hat{f}_n\left(\mathbf{X}^*\right)\right)^2\right]$$

  is very difficult to estimate

- Something similar:

$$E_{\{(Y^*,\mathbf{x}^*),\text{data}\}}\left[\left(Y^* - \hat{f}_n\left(\mathbf{X}^*\right)\right)^2\right]$$

  is easier to estimate
  (what is the difference, exactly?)

Discussion points

- Goodness of fit vs. prediction power
- How do we estimate prediction MSE?

$$E_{(Y^*,\mathbf{X}^*)}\left[\left(Y^* - \hat{f}\left(\mathbf{X}^*\right)\right)^2\right]$$

- Can it be done without a test set?

Mean squared prediction error

- How do we estimate the MSPE

$$E_{(Y^*,\mathbf{x}^*)}\left[\left(Y^* - \hat{f}\left(\mathbf{X}^*\right)\right)^2\right] \quad ?$$

- (a): “Training / Testing” data sets
- (b): “Recycling” a single data set
- (c): “Direct estimates”

Test set approach

- Randomly split the data into a training set and a test set
- “Train” (“estimate”, “fit”) your model(s) using only the training set
- Use the estimated model(s) to obtain predictions for the test set (only)

Test set approach

- Check how well your model(s) did

$$\widehat{\text{MSPE}} = \frac{1}{n_T} \sum_{j \in \mathcal{T}} (y_j - \hat{y}_j)^2$$

where:

$\mathcal{T}$ is the **test set**,

n_T is the size of $\mathcal{T}$, and

$\hat{y}_j$ are the **predicted values**

Predictions

- Back to the Pollution example
- Read the training set
- Train the full and the reduced models
- Read the test set
- Use both models to predict `MORT`
- Compare both sets of predictions

Predictions

```
> x.te <- read.table('pollution-test.dat',...
>
> x.te$pr.full <- predict(full, newdata=x.te)
> x.te$pr.reduced <- predict(reduced,
        newdata=x.te)
>
> with(x.te, mean( (MORT - pr.full)^2 ))
[1] 4677.45
>
> with(x.te, mean( (MORT - pr.reduced)^2 ))
[1] 1401.571
```

Discuss

Predictions

- Back to the Pollution example
- Repeat with a different training / test split
- Compare conclusions

Test set approach

- **Pros**:
- Estimates what we (generally) want to estimate
- It is computationally very fast

Test set approach

- **Cons**:
- Estimated MSPE depends on the training set
- Estimated MSPE may be quite variable - why is this a problem?
- One does not use all the available information for training the model(s)

Cross validation

- Leave-one-out
- Description
- Picture

CV - Pollution data

- Run it on the pollution data

CV - leave-one-out

- **Pros**:
- It is not random!
- It uses almost all of the data for training
- **Cons**:
- It can be numerically very expensive (although “shortcuts” for linear models exist).

Cross validation

- K-fold
- Description
- Picture

CV - Revisit example data

- Run it on the pollution example

K-fold CV

- **Pros**:
- Faster than leave-one-out
- Less variable estimator for MSPE than leave-one-out (explain)
- Training set is larger than "training/test" method - less biased MSPE estimate (but more biased than leave-one-out)

K-fold CV

- **Cons**:
- Choosing K is not trivial (K = 5 or K = 10 have been shown empirically to work well)
- Training set is smaller than leave-one-out's
- It is a random estimate of MSPE (it is a good idea to run it several times)

Discussion points

- Why compare models? Why not choose the “largest” (best fit to the data)?

Discussion points

- Correlated covariates:
- May reduce prediction accuracy
- Mask each other when included simultaneously in a model

Discussion points

- Correlated covariates have become prevalent
- Researchers can (and do) collect data “blindly”
- Data are collected without a specific question in mind