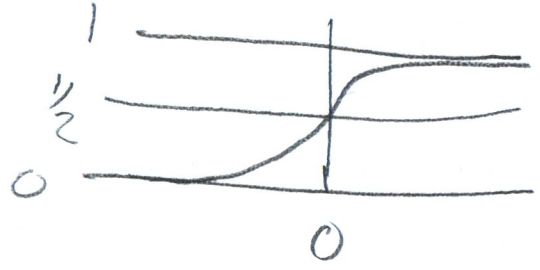
Linear Model

$[-\infty, \infty]$ $y = \beta_0 + \beta_1 x_1 + \cdots$

$$\underbrace{\log(\text{odds}(y=1))}_{t} = \beta_0 + \beta_1 x_1 + \cdots$$

$[0, 1]$ $y = \dfrac{1}{1+e^{-t}}$



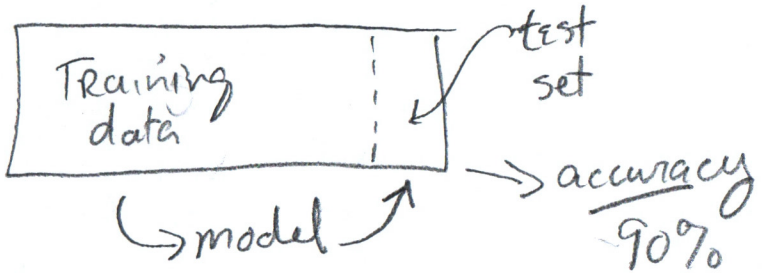
{ Thou shall not test on training data. }

LEARNING w/ MEMORIZING

↕

Generalization : the ability to correctly predict new conditions

1. Holdout.



Holdout (BIG)

14.3 Cross Validation

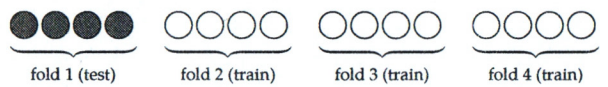
k-fold (small-medium)

Cross validation is an alternative to holdout. In cross validation, all points in the dataset are used for training and testing, but never at the same time. Cross validation begins by splitting the dataset into a set of k groups of roughly equal size. Each group of data is called a *fold*, and points are randomly assigned to the folds. For example, at dataset with 16 points could be divided into $k = 4$ folds.

To begin cross validation, one of the folds is set aside for validation, similar to holdout. A model is trained using the remaining $k - 1$ folds and tested against the holdout fold.

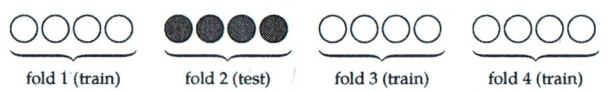
5/4

3/4



Next we put fold 1 back into the training set and set aside fold 2 for testing. Then we re-train our model using folds 1, 3, and 4 and validate with fold 2.

2/4



This process continues $k = 4$ times, with the final model trained on folds 1–3 and validated with the data in fold 4. The final step is to average the accuracies across all k folds. This average is reported as the final accuracy, and a full model can be trained using all of the data.

The advantage of k-fold cross validation is that every point in the dataset is used for testing, so the method is not sensitive to which data are selected for holdout. However, the method is still stochastic as the accuracy of each model depends on how the data are randomly assigned to the folds. A k-fold cross validation requires training k separate models in addition to the final model with all of the data. This might be costly for very large datasets, so cross validation is more common in small- to medium-sized problems.

14.3.1 Leave-one-out Cross Validation

There is no rule for determining the number of folds (k) for a cross validation. Smaller datasets benefit from higher values of k since fewer points are held out

(small)

Leave-one-out C.V.

k = # of points (n)

- most computation
  n models
- best measure of accuracy.

ERROR = BIAS + VARIANCE

Consistently under-fitting data

overfitting data inconsistency