Classification: The response variable is categorical

Example: Y_i = party affiliation of a survey respondent ("Dem", "Ind", or "Rep")

sometimes assign #'s → 1, 2, 3

x_i = age of a survey respondent in years

Two possible goals:

1) Estimate class membership probability functions (1 per class)

$f_1(x_i)$ = probability that a person with age x_i is a "Dem"

$f_2(x_i)$ = " " " " " " " " "Ind"

$f_3(x_i)$ = " " " " " " " " "Rep"

Note probabilities have to add up to 1 for each x:

$$f_1(x) + f_2(x) + f_3(x) = 1$$

2) Estimate the most likely class membership for a person with covariate x_i:

$\hat{y}_i = k$ if $f_k(x_i) > f_j(x_i)$ for each other category j.

Example: Suppose we estimate

$f_1(50) = 0.47$, $f_2(50) = 0.7$, $f_3(50) = 0.46$

$\hat{y}_i = 1$, i.e. we guess that a 50 year old is a "Dem".

Ties are usually broken at random.

How Good Are Our Predictions?

Many ways to measure. One way is:

Classification Error Rate: The proportion of test set observations for which our predicted class is wrong:

$$\frac{\#\text{ incorrect predictions}}{\#\text{ obs. in test set}} = \frac{1}{n}\underbrace{\sum_{i=1}^{n}}_{\text{sum over test set}} \mathbb{I}(\hat{y}_i \neq y_i)$$

Where $\mathbb{I}(\cdot)$ is the indicator function:

$$\mathbb{I}(\hat{y}_i \neq y_i) = \begin{cases} 1 & \text{if } \hat{y}_i \neq y_i \\ 0 & \text{if } \hat{y}_i = y_i \end{cases}$$

Basically $\sum_{i=1}^{n} \mathbb{I}(\hat{y}_i \neq y_i)$ is a count of how many predictions were wrong.

Cross-validation

All that changes is evaluation by classification error rate instead of MSE.

1. Obtain observation indices for k folds
2. Allocate space to store validation set error rates
3. For each fold $i = 1, \ldots, k$:
   a. Create validation set using fold i, train set using the other folds put together
   b. fit model to train set
   c. get predictions for validation set
   d. calculate validation set error rate
   e. save validation set error rate in space from step 2.

KNN for Classification

For class j,

$\hat{f}_j(x_0)$ = proportion of the neighbors of x_0 that are in class j.

$$= \frac{1}{k} \sum_{i \in N_0^{(k)}} \mathbb{I}(y_i = j)$$

sum over observations from training set in neighborhood of x_0.

Example: Suppose we have $n = 8$ observations:

age (x_i)	party (y_i)
10	Ind
20	Ind
30	Dem
40	Rep
50	Dem
60	Dem
70	Rep
80	Rep

Using KNN with $K=3$, what are:

~~What are~~

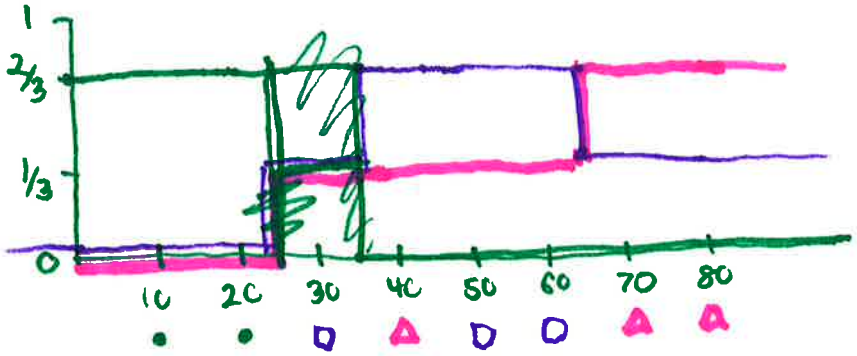
$\hat{f}_1(47) = \frac{1}{3}(0 + 1 + 1) = \frac{2}{3}$

$\hat{f}_2(47) = \frac{1}{3}(0 + 0 + 0) = 0$

$\hat{f}_3(47) = \frac{1}{3}(1 + 0 + 0) = \frac{1}{3}$

$\hat{y}_i = 1$ ("Dem")

The full class membership probability functions:



- $\hat{f}_2(x)$ = Estimated prob. - Independent