Data Mining with Weka

Logistic regression

Ian H. Witten

Logistic regression

Can do better by using prediction probabilities

Probabilities are often useful anyway …

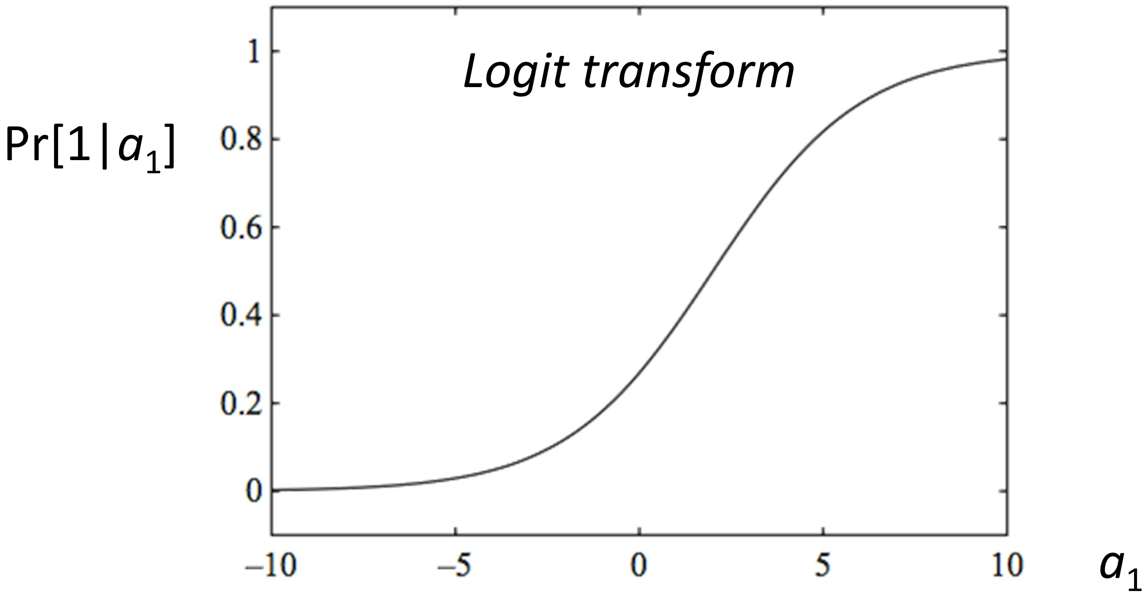
- Naïve Bayes produces them (obviously)
  - *Open diabetes.arff and run Bayes>NaiveBayes* with 90% percentage split
  - *Look at columns: actual, predicted, error, prob distribution*
- Other methods produce them too …
  - *Run rules>ZeroR. Why probabilities* [*0.648, 0.352*] for [*tested_negative, tested_positive*]?
  - *90% training fold has 448 negatve, 243 positive instances*
  - *(448+1)/(448+1 + 243+1) = 0.648* [*cf. Laplace correction, in the* Simplicity first *video*]
  - *Run trees>J48*
  - *J48 uses probabilities internally to help with pruning*

Make linear regression produce probabilities too!

Logistic regression

- Linear regression: calculate a linear function and then a threshold
- Logistic regression: estimate class probabilities directly

$$\Pr[1 \mid a_1, a_2, \ldots, a_k] = 1/(1+\exp(-w_0 - w_1 a_1 - \ldots - w_k a_k))$$



- Choose weights to maximize the log-likelihood (not minimize the squared error):

$$\sum_{i=1}^{n} (1-x^{(i)})\log(1-\Pr[1 \mid a_1^{(1)}, a_2^{(2)}, \ldots, a_k^{(k)}]) + x^{(i)}\log(\Pr[1 \mid a_1^{(1)}, a_2^{(2)}, \ldots, a_k^{(k)}])$$

Logistic regression

- Open file diabetes.arff

Classification-by-regression		76.8%	mean of 10 runs
cf	ZeroR	65.1%	65.1%
	Naïve Bayes	76.3%	75.8%
	J48	73.8%	74.5%
Apply functions>Logistic		77.2%	77.5%

- Extension to multiple classes …
  - Perform a regression for each class?
    (like multi-response regression)
  - No. Probabilities won't sum to 1
  - Can be tackled as a joint optimization problem

Logistic regression

- Logistic regression is popular and powerful
- Uses logit transform to predict probabilities directly
  - like Naïve Bayes
- Also learned about
  - Prediction probabilities from other methods
  - How to calculate probabilities from ZeroR