Learning from Data: Logistic Regression

Amos Storkey, School of Informatics

October 20, 2005

http://www.anc.ed.ac.uk/∼amos/lfd/

Recap

- Classification problems:
- On the basis of historical information, classify a new instance as belonging to a particular class.
  - Training data with targets $(\mathbf{x}, t)$.
  - Sometimes validation data with targets.
  - Test data: targets are only visible for evaluation of method.
- Have used class conditional modelling: $P(t|\mathbf{x}) \propto P(\mathbf{x}|t)P(t)$. This is a *generative* approach.
- Now model $P(t|\mathbf{x})$ directly. This is a *discriminative* approach. Don't bother modelling $P(\mathbf{x})$.

Which is the correct model?

- Two approaches encode different assumptions.
- Generative assumption: classes exist because data is drawn from two different distributions.
- Discriminative assumption, class label is drawn dependent on the value of $\mathbf{x}$.
- Generative: Class $\rightarrow$ Data.
- Discriminative: Data $\rightarrow$ Class.

Example

- The weight of men and women. Men and women have different weight distributions because of characteristics of gender: men are on average taller, and are therefore more likely to have a higher weight.
- Weight and heart attacks. Obesity is a contributory factor to heart attacks. We do not expect someone's current weight to be determined by the heart attack they are going to have in the future!
- The underlying distribution of people's weight does affect the chance of someone with a given weight having a heart attack. E.g. if the whole population on average lost weight, does not affect the model.
- Can ignore the distribution of people's weight.

Is this rule hard and fast?

- No. In a given stationary (i.e. no distributions are changing) circumstance, with no missing data, either approach can be used.
- If the discriminative approach is used in a situation where a generative approach is more appropriate, it just models the $P(\mathbf{x}|t)$ and $P(t)$ implicitly through $P(t|\mathbf{x}) = P(\mathbf{x}|t)P(t)/P(\mathbf{x})$.
- The discriminative approach often has the advantage that more flexible model can be used for $P(t|\mathbf{x})$ than for $P(\mathbf{x}|t)$.

PMR versus LfD

- This is where PMR and LfD diverge.
- PMR is more to do with generative modelling, especially through the use of belief networks.
- LfD is going to focus on discriminative modelling, especially through neural networks and related methods.

Two Class Discrimination

- Consider a two class case: $t \in \{0, 1\}$.
- Use a model of the form

$$P(t = 1|\mathbf{x}) = f(\mathbf{x}; \mathbf{w})$$

- f must be between 0 and 1. Furthermore the fact that probabilities sum to one means

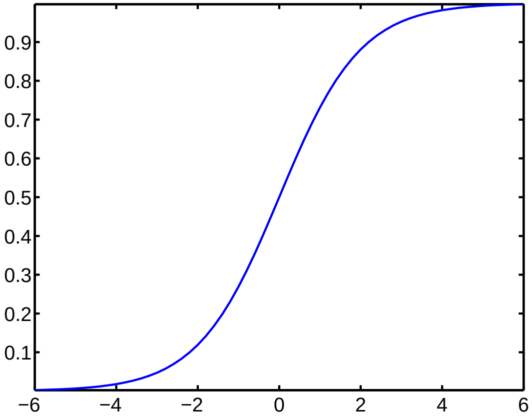
$$P(t = 0|\mathbf{x}) = 1 - f(\mathbf{x}; \mathbf{w})$$

- What form should we use for f?

The logistic function

- We need two things:
- A function that returns probabilities (i.e. stays between 0 and 1).
- A means of incorporating $\mathbf{x}$ dependencies through the parameters $\mathbf{w}$.
- The logistic (or sigmoid) function provides the first of these.
- $f(x) = \sigma(x) \equiv 1/(1 + \exp(-x))$.
- As x goes from $-\infty$ to ∞, so f goes from 0 to 1.

The Logistic Function

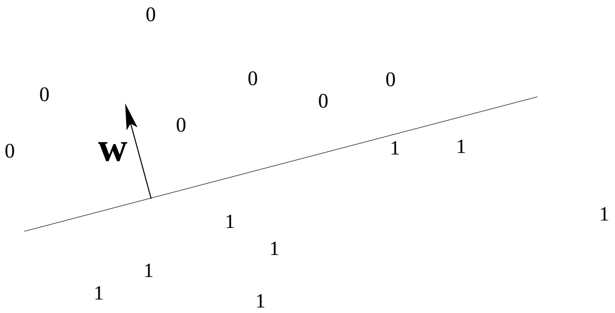


The Logistic Function $\sigma(x) = \frac{1}{1+\exp(-x)}$.

The linear weights

- We need two things:
- A function that returns probabilities (i.e. stays between 0 and 1).
- A means of incorporating $\mathbf{x}$ dependencies through the parameters $\mathbf{w}$.
- A linear weighting scheme provides the second of these:
- $P(t = 1|\mathbf{x}) = \sigma(b + \mathbf{x}^T\mathbf{w})$.
- $\sigma(x) = 0.5$ when $x = 0$. Hence the decision boundary is given by $\mathbf{x}^T\mathbf{w} = -b$.
- Decision boundary is a $d - 1$ hyperplane for a d dimensional problem.

The Linear Decision Boundary



For two dimensional data the decision boundary is a line.

Logistic regression

- The bias parameter b shifts the position of the hyperplane, but does not alter the angle.
- The direction of the vector $\mathbf{w}$ affects the angle of the hyperplane. The hyperplane is perpendicular to $\mathbf{w}$.
- The magnitude of the vector $\mathbf{w}$ effects how certain the classifications are.
- For small $\mathbf{w}$ most of the probabilities within a region of the decision boundary will be near to 0.5.
- For large $\mathbf{w}$ probabilities in the same region will be close to 1 or 0.

The Perceptron

- The perceptron is the special case of logistic regression where the magnitude of **w** tends to infinity.
- Absolutely certain classification: all probabilities are 0 or 1.
- Define $\theta(x) = 1$ if $x > 0$ and 0 otherwise.
- Have $p(c = 1|x) = \theta(b + \mathbf{x}^T\mathbf{w})$.

Learning Logistic Regressors

- Want to set $\mathbf{w}$ and b using training data.
- As before:
  - Write out the model and hence the likelihood
  - Find the derivatives of the log likelihood w.r.t the parameters.
  - Adjust the parameters to maximize the log likelihood.

Likelihood

- Assume data is independent and identically distributed.
- The likelihood is

$$p(D) = \prod_{i=1}^{N} P(t^i|\mathbf{x}^i) = \prod_{i=1}^{N} P(t=1|\mathbf{x}^i)^{t^i} \left(1 - P(t=1|\mathbf{x}^i)\right)^{1-t^i} \quad (1)$$

- Hence the log likelihood is

$$\log P(D) = \sum_{i=1}^{N} t^i \log P(t=1|\mathbf{x}^i) + (1-t^i) \log \left(1 - P(t=1|\mathbf{x}^i)\right) \quad (2)$$

Logistic Regression Log Likelihood

- Using our assumed logistic regression model, the log likelihood becomes

$$\log P(D|\mathbf{w}, b) = \sum_{i=1}^{N} t^i \log \sigma(b+\mathbf{w}^T\mathbf{x}^i)+(1-t^i)\log\left(1-\sigma(b+\mathbf{w}^T\mathbf{x}^i)\right) \tag{3}$$

- We wish to maximise this value w.r.t the parameters **w** and *b*.
- Cannot do this explicitly as before. Use an iterative procedure.
- This will be considered in the next lecture.

Summary

- The difference between generative and discriminative models.
- The logistic function.
- Logistic regression.
- Hyperplane decision boundaries.
- The Perceptron.
- The likelihood for logistic regression.