COMP90051 **Statistical Machine Learning**

Semester 2, 2015

Lecturer: Ben Rubinstein

3. Linear Approaches

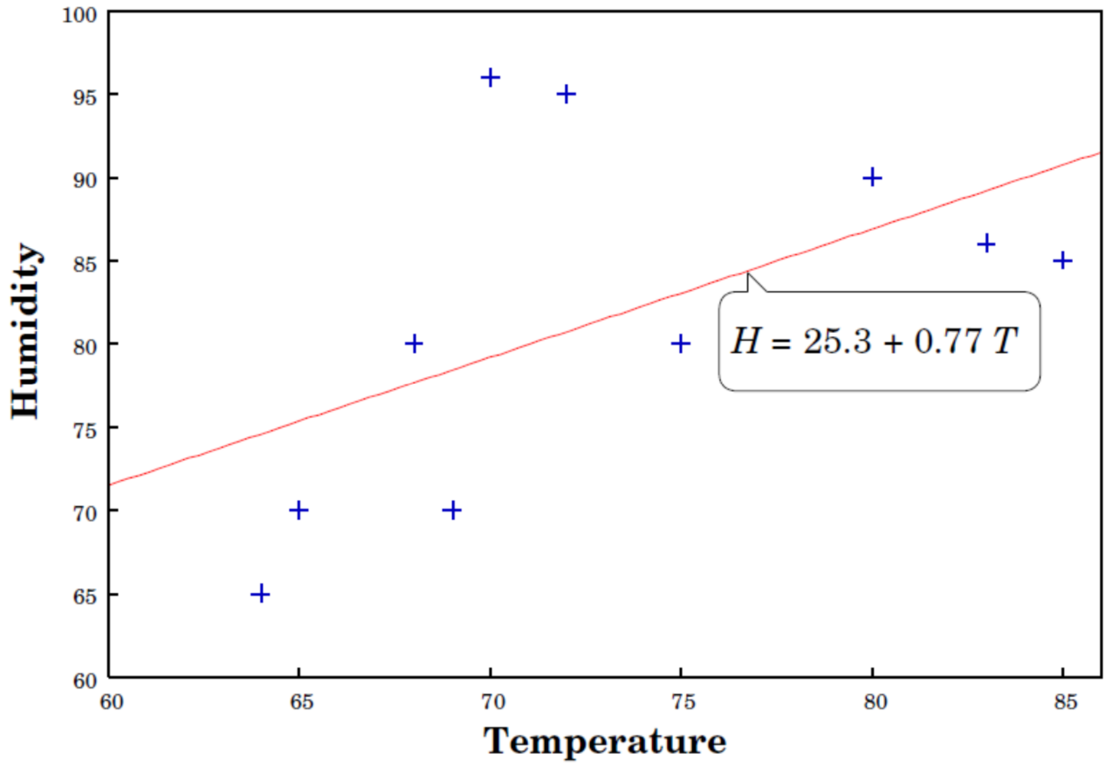
THE UNIVERSITY OF MELBOURNE

Linear Approaches

These need less data, are interpretable and often work well in practice. They also serve as the basis for many non-linear approaches

Example: Predict Humidity from Temperature

Temperature	Humidity
TRAINING DATA	
85	85
80	90
83	86
70	96
68	80
65	70
64	65
72	95
69	70
75	80
TEST DATA	
75	70



In regression, response ("class") numeric

Linear regression: H=a+bT

Example (cont.)

- Fitting linear regression H=a+bT via least square errors
  - a,b to minimise $\sum_{i=1}^{10}\left(H_i - \widehat{H}_i\right)^2 = \sum_{i=1}^{10}(H_i - a - b\,T_i)^2$
  - Set derivatives to zero

$$\frac{\partial}{\partial a}\text{ gives: } 0 = \sum_{i=1}^{10}(H_i - a - b\,T_i)$$

$$\frac{\partial}{\partial b}\text{ gives: } 0 = \sum_{i=1}^{10} T_i(H_i - a - b\,T_i)$$

  - Rewrite as system of linear equations

$$\begin{pmatrix} 10 & \sum_{i=1}^{10} T_i \\ \sum_{i=1}^{10} T_i & \sum_{i=1}^{10} {T_i}^2 \end{pmatrix}\begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{10} H_i \\ \sum_{i=1}^{10} T_i H_i \end{pmatrix}$$

  - Solution gives linear regression! a=25.3, b=0.77

Take-away point:
Main step intuition:
- Write derivative
- Set to zero
- Solve for model

Linear Regression

- Assume linear relationship between attributes, response

$$y \approx w_0 + \sum_{j=1}^{d} w_j x_j$$

for response y, instance vector* $\mathbf{x}$, weight vector $\mathbf{w}$

- **Trick:** augment instance with $x_0 = 1$

$$y \approx \sum_{j=0}^{d} w_j x_j = \mathbf{x}'\mathbf{w}$$

* **Bold** denotes vector (when lowercase) or matrix (when uppercase); prime ‘ denotes transpose

Method of Least Squares

- Method of least squares
  - For convenience, roll up data into *n* by *d* matrix $\mathbf{X}$, and *n* by 1 $\mathbf{y}$
  - Choose $\mathbf{w}$ to minimise residual sum of squares *

    $RSS(\mathbf{w}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \mathbf{X}_{i}.\mathbf{w})^2$
  - Set derivative to 0 and solve for $\mathbf{w}$ yields

    $\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ when inverse exists (the "normal equations")

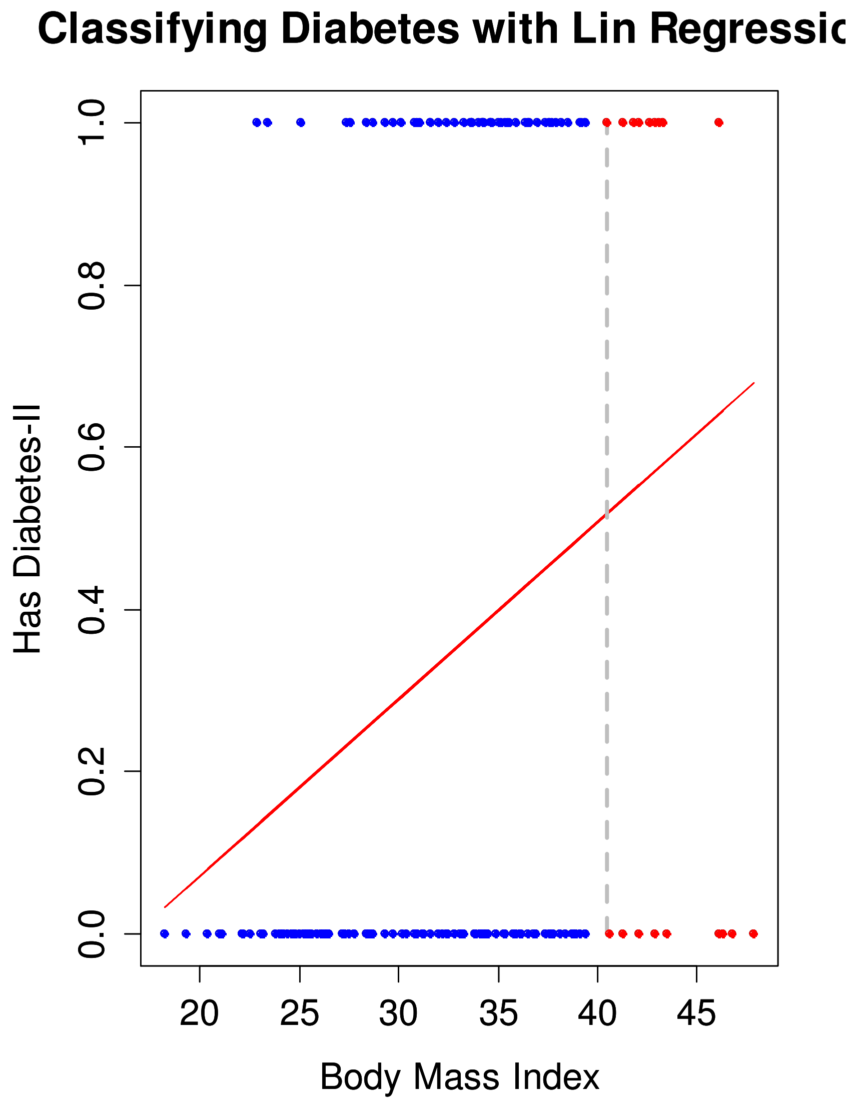
Take-away point:
Main step intuition:
- Write derivative
- Set to zero
- Solve for model

- Decision-theoretic interpretation (discriminative)
  - Minimises risk under square loss (NB: here risk=bias^2+var)
- Statistical model (discriminative)
  - MLE when $\Pr(Y|X) = X'\mathbf{w} + Normal(0, \sigma^2)$

* The notation $\mathbf{X}_{i}.$ means the i^{th} row of matrix $\mathbf{X}$, which is a row vector.

Linear Approaches to Classification

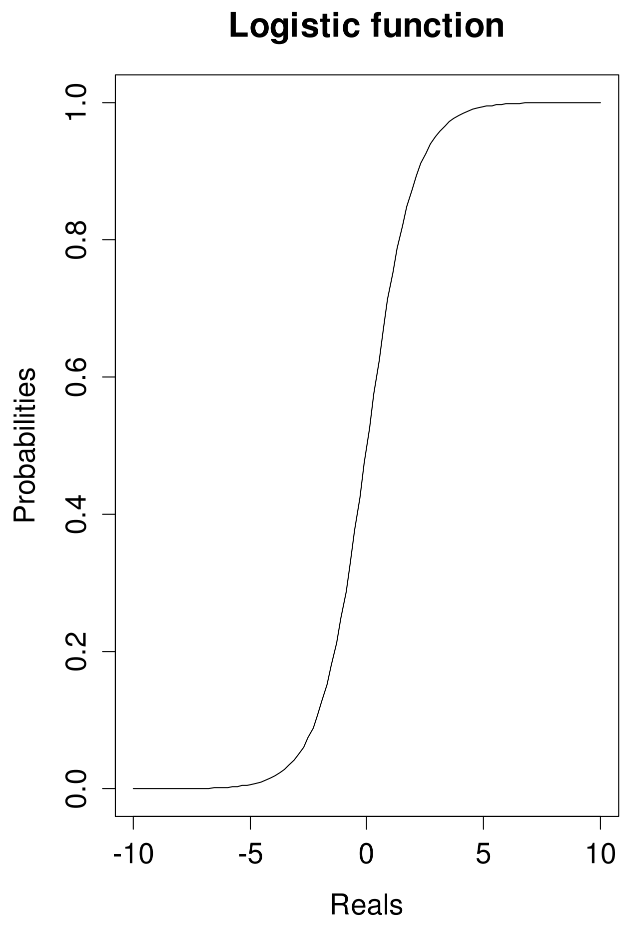
- Problem: given BMI does a patient have T2D?
- Have already seen one!
  - Linear regression
- Will see more later in subject
  - Naïve Bayes
  - SVM with linear kernel



Logistic Regression

- Probabilistic classification
  - $\Pr(Y = true|X = x) = f(x)$
  - Could we use linear regression? $f(\mathbf{x}) = \mathbf{x}'\mathbf{w}$
- Problem: LHS in [0,1], RHS arbitrary real
- So use: $logistic(x) = \frac{1}{1+\exp(-x)}$
  - $\Pr(Y = true|X = \mathbf{x}) = logistic(\mathbf{x}'\mathbf{w})$
  - Equivalent to linear model for "log-odds"

$$\log \frac{\Pr(Y = true|X = \mathbf{x})}{\Pr(Y = false|X = \mathbf{x})} \approx \mathbf{x}'\mathbf{w}$$



Logistic Regression (cont.)

- Training is again via MLE
  - Results in choosing weights by

$$\arg\max_{\mathbf{w}} \prod_{i=1}^{n} p(y_i|\mathbf{x}_i)$$

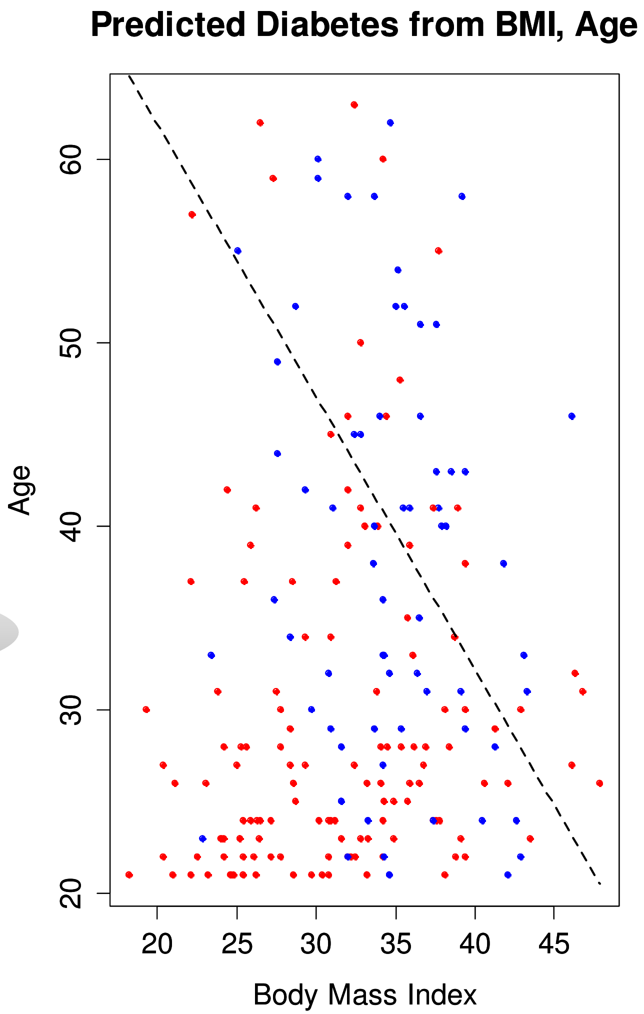
log trick

$$= \arg\max_{\mathbf{w}} \textstyle\sum_{i=1}^{n} \log p(y_i|\mathbf{X}_{i\cdot})$$

$$= \arg\max_{\mathbf{w}} \sum_{i=1}^{n} \left(y_i \mathbf{X}_{i\cdot}\mathbf{w} - \log\left(1 + e^{\mathbf{X}_{i\cdot}\mathbf{w}}\right)\right)$$

using: *y* is 0 or 1

  - Setting derivative to 0 yields "score equations"; Solution is done **iteratively** via Newton-Raphson algorithm (HTF pg. 120)
  - Called "Iteratively-reweighted least squares"



Summary

- Linear supervised approaches
  - For regression: Linear regression
  - For classification: Logistic regression