- Probabilistic Interpretation Of Linear Regression
- Feature Selection
- Maximum likelihood vs MAP estimation

Recap: Probabilistic Interpretation Of Logistic Regression

- Observation Y is a random variable. $P(Y = y|\theta, x) = \frac{1}{1+e^{-y\theta^T x}}$ with y=+1 or -1. Each observation is independent.
- So for the i'th training data point, $P(Y = y^{(i)}|\theta, x^{(i)}) = \frac{1}{1+e^{-y^{(i)}\theta^T x^{(i)}}}$
- And the **likelihood** $P(d|\theta)$ of the observing all m training data points d is:

$$P(d|\theta) = \prod_{i=1}^{m} \frac{1}{1 + e^{-y^{(i)}\theta^T x^{(i)}}}$$

  (we multiply because the observations are independent)
- Taking logs, the **log-likelihood** is
  $\sum_{i=1}^{m} \log \frac{1}{1+e^{-y^{(i)}\theta^T x^{(i)}}} = -\sum_{i=1}^{m} \log(1 + e^{-y^{(i)}\theta^T x^{(i)}})$
  (recall $\log \frac{1}{z} = -\log z$)
- Select parameter value θ that *maximises* the log-likelihood (same as maximising the likelihood). Same as selecting θ that *minimises* -loglikelihood i.e. minimises cost $\sum_{i=1}^{m} \log(1 + e^{-y^{(i)}\theta^T x^{(i)}})$.

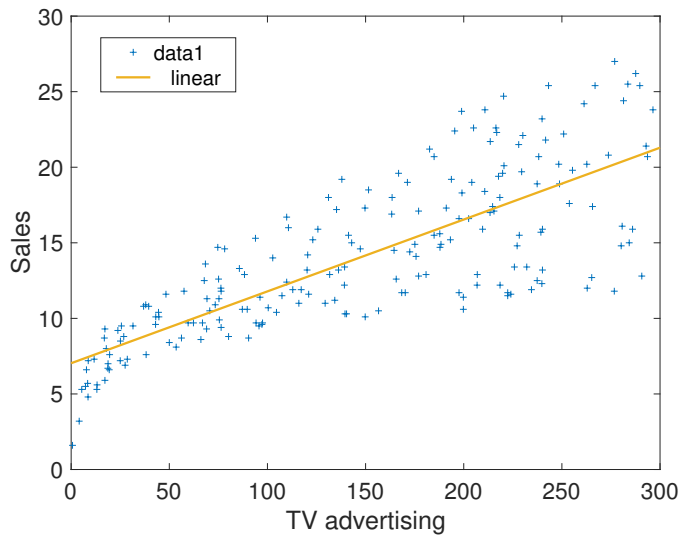
General Maximum Likelihood Approach

- Define statistical model of observations,
  e.g. $P(Y = y|\theta, x) = \frac{1}{1+e^{-y\theta^T x}}$, y=+1 or -1, observations independent
- Write expression for likelihood,
  e.g. $P(d|\theta) = \prod_{i=1}^{m} \frac{1}{1+e^{-y^{(i)}\theta^T x^{(i)}}}$
- Select parameter value θ that minimises -loglikelihood $-\log P(d|\theta)$ (which is the same as maximising the likelihood $P(d|\theta)$)
- Note: scaling factors such as $1/m$ in the log-likelihood don't change the θ value that minimises the function. E.g. $\sum_{i=1}^{m} \log(1 + e^{-y^{(i)}\theta^T x^{(i)}})$ and $\frac{1}{m}\sum_{i=1}^{m} \log(1 + e^{-y^{(i)}\theta^T x^{(i)}})$ are both minimised by the same value of θ.

Linear Regression

Recall linear regression setup:

- Hypothesis: $h_\theta(x) = \theta^T x$
- Parameters: θ
- Cost Function: $J(\theta) = J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$
- Learn Parameters: Select θ that minimises $J(\theta)$. E.g. can find θ using gradient descent.



Probabilistic Interpretation: Linear Regression

- Statistical model: output Y is a Gaussian random variable with mean $\theta^T x$ and variance 1 i.e. its PDF is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\theta^T x)^2}{2}}$$

- The likelihood $f_{D|\Theta}(d|\theta)$ of the training data d is therefore:

$$f_{D|\Theta}(d|\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}} e^{-\frac{(y^{(i)}-h_\theta(x^{(i)})^2}{2}}$$

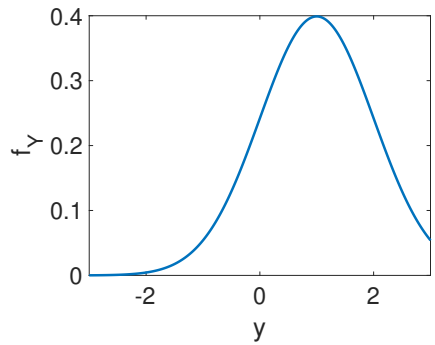
- Taking logs: $\log f_{D|\Theta}(d|\theta) = -\log(\frac{1}{\sqrt{2\pi}^m}) \sum_{i=1}^{m} \frac{(y^{(i)}-h_\theta(x^{(i)})^2}{2}$
- Choose parameter value θ that minimises $-\log f_{D|\Theta}(d|\theta) = \log(\frac{1}{2\sqrt{2\pi}^m}) \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})^2$. Same as choosing θ that minimises $\frac{1}{m} \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})^2$

A Note on Notation

- Output Y is a Gaussian random variable with mean $\theta^T x$ and variance 1 i.e. its PDF is

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\theta^T x)^2}{2}}$$

- Often also written as $Y = \theta^T x + M$ where $M \sim N(0, 1)$ i.e. M is Gaussian "noise" with mean 0 and variance 1.
  - Y is Gaussian. Aadding a constant just shifts the PDF, but its remains Gaussian, e.g. adding -1 gives:

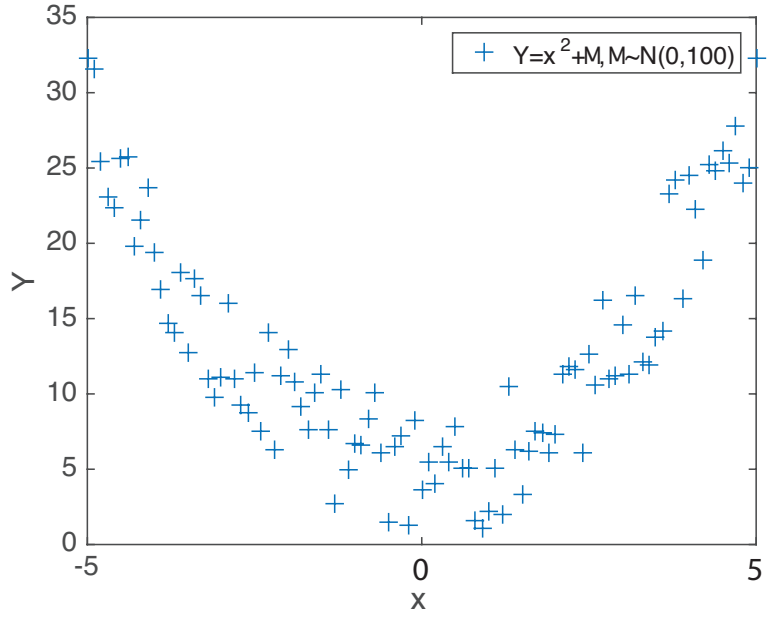


  - $E[Y] = E[\theta^T x + M] = E[\theta^T x] + E[M] = \theta^T x$ since $E[M] = 0$
  - $Var(Y) = Var(\theta^T x + M) = Var(M) = 1$ (adding a constant to a RV doesn't change its variance)

Fitting nonlinear curves: choosing features

Same as with logistic regression we can modify the features to allow fitting of nonlinear curves. E.g.

- Suppose output is quadratic $\theta_0 + \theta_1 x^2$

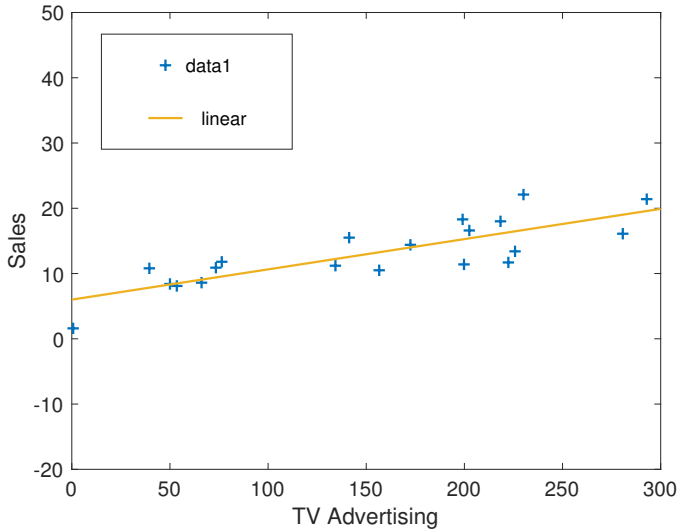


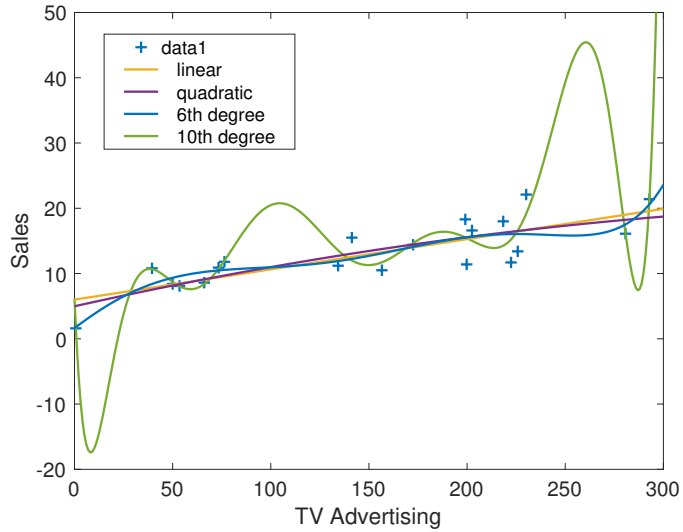
- Define feature $z = x^2$ and use hypothesis $\theta_0 + \theta_1 z$. Now can directly apply linear regression.

Feature Selection

Advertising example again. Thin out data by taking every 10th point. Try a few different hypothesis:

- $h_\theta(x) = \theta_0 + \theta_1 x$
- $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$
- $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_6 x^6$
- $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_{10} x^{10}$

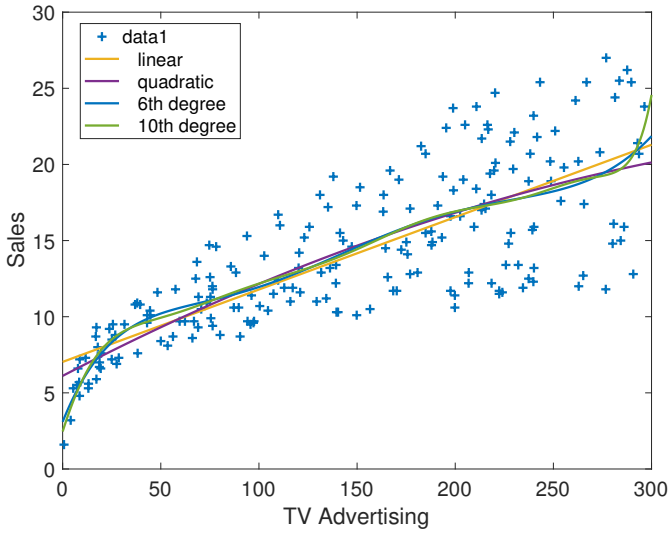


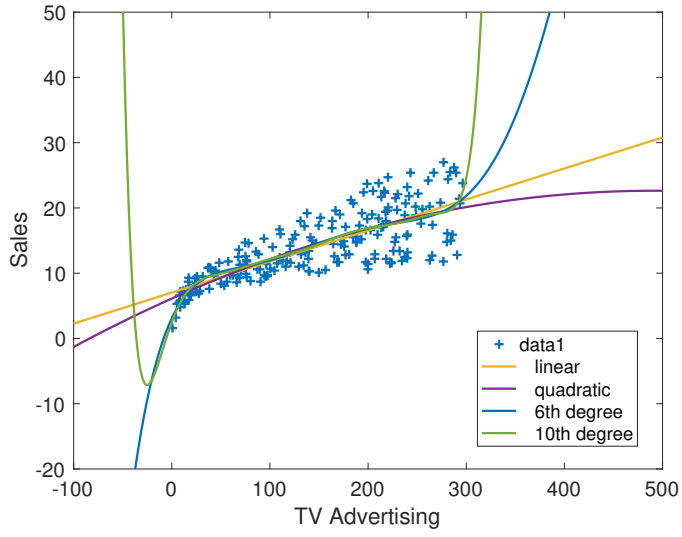


- As we add more parameters, we start to fit the "noise" in the training data, called **overfitting**.
- But of use too few parameters then will get a poor fit, **underfitting**.
- How to strike the right balance between these ? This is an example of the **bias-variance trade-off**.

Feature Selection

More data can help, e.g. when don't thin out data:





- But even with more data, still our hypothesis doesn't generalise well i.e. doesn't predict well for data outside the training set.

Feature Selection

There are two main approaches to feature selection (both can be used together):

- **Sequential Model Selection**:
  repeat {Add a new feature, fit model, calc cost func} until change in cost function is small.
- **Regularisation**:
  Change the cost function to add a penalty e.g. in linear regression change cost fn to $\frac{1}{m}\sum_{i=1}^{m}(\theta^T x^{(i)} - y^{(i)})^2 + \frac{1}{\lambda}\sum_{j=1}^{n}\theta_j^2$

  Here $\frac{1}{\lambda}\sum_{j=1}^{n}\theta_j^2$ is the **penalty** term, decreasing λ makes this term bigger, increasing λ makes it smaller (when $\lambda = \infty$ then $\frac{1}{\lambda} = 0$ and we are back to original setup with no penalty).

Probabilistic Interpretation of Regularisation

Recall Bayes Rule for PDFs

$$f_{\Theta|D}(\vec{\theta}|d) = \frac{f_{D|\Theta}(d|\vec{\theta}) f_{\Theta}(\vec{\theta})}{f_D(d)}$$

posterior *likelihood* *prior*

- **Maximum likelihood estimate** of θ maximises the likelihood $f_{D|\Theta}$ (equivalently, minimises $-\log f_{D|\Theta}$).
- **MAP** (maximum a posteriori) estimate of θ maximises the posterior.
  - Taking logs, $\log f_{\Theta|D} = \log f_{D|\Theta} + \log f_{\Theta} - \log f_D$.
  - Now $\log f_D$ is just a normalising constant that doesn't depend on θ, so the MAP estimate maximises $\log f_{D|\Theta} + \log f_{\Theta}$ or, equivalently, minimises $-\log f_{D|\Theta} - \log f_{\Theta}$
  - $-\log f_{\Theta}$ (i.e. the log-prior) acts like the penalty term when using regularisation.

Probabilistic Interpretation of Regularisation

Example: Ridge Regression (a variant of Linear Regression)

- Statistical model: output Y is Gaussian with mean $\theta^T x$ and variance 1 i.e. $-\log f_{\Theta|D}(\theta|d) \propto \sum_{j=1}^{m}(y^{(j)} - \theta^T x_i^{(j)})^2$.
- Notation: $\propto$ reads "is proportional to" i.e. ignoring scaling factors such as $\log(1/\sqrt{2\pi}^m)$ since they don't change the θ value that minimises the function.
- Assume prior is that each element θ_i of parameter vector θ is Gaussian with mean 0 and variance $\lambda/2$ i.e. $f_{\Theta_i}(\theta_i) \propto \exp(-\frac{\theta_i^2}{\lambda})$ and so $-\log f_{\Theta_i}(\theta_i) \propto \frac{\theta_i^2}{\lambda}$. Also assume that the θ_i's are independent.
- Then $f_\Theta(\theta) = \prod_{i=1}^{n} f_{\Theta_i}(\theta_i)$ and $\log f_\Theta(\theta) = \sum_{i=1}^{n} \log f_{\Theta_i}(\theta_i) \propto -\sum_{i=1}^{m} \frac{\theta^2}{\lambda}$.
- Combining these using Bayes Rule gives:

$$-\log f_{\Theta|D}(\theta|d) \propto \sum_{j=1}^{m}(y^{(j)} - \theta^T x_i^{(j)})^2 + \frac{1}{\lambda}\sum_{i=1}^{n}\theta_i^2$$

which is the cost function for linear regression modified by addition of the penalty $\frac{1}{\lambda}\sum_{i=1}^{n}\theta_i^2$