TRINITY COLLGE DUBLIN

School of Computer Science and Statistics

Extra Questions ST3009: Statistical Methods for Computer Science

Question 1. (a) Give Bayes rule for PDFs

(b) Explain the difference between the maximum likelihood and the MAP estimate of a random variable

(c) Suppose after observing data the likelihood of parameter θ is $L(\theta) = e^{-(\theta-1)^2}$. What is the maximum likelihood estimate of θ ?

Solution

- The value of θ which maximises $e^{-(\theta-1)^2}$ is $\theta = 1$

Question 2. Suppose and urn contains balls and that fraction θ of the balls are white and the rest are red. I draw n balls, with replacement, from the urn and let X be the number of white balls observed.

(a) Give an expression for the likelihood $P(X = x|\theta)$

(b) Suppose $n = 100$ and I observe 25 white balls. What is the maximum likelihood estimate for θ (use matlab to plot the value of $P(X = x|\theta)$ for a range of values of θ).

(c) Suppose now that before drawing the balls my prior probability was $P(\theta) = \frac{1}{20\pi}e^{-100(\theta-0.5)^2}$ and for simplicity assume that $P(X = 25) = 1$ (since it just scales the posterior). Give an expression for the posterior $P(\theta|X = x)$ (use Bayes rule).

(d) What is the MAP estimate for θ (use matlab to plot the value of $P(\theta|X = x)$ for a range of values of θ). Discuss why it differs from the maximum likelihood estimate.

Solution

- The probability of drawing x white balls is $P(X = x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$.
- The maximum likelihood estimate is $\theta = 0.25$
- The posterior is $P(\theta|X = x) = P(X = x|\theta)P(\theta)/P(X = x) = \frac{1}{20\pi}\binom{n}{x}\theta^x(1-\theta)^{n-x}e^{-100(\theta-0.5)^2}$.
- The MAP estimate is approximate $\theta = 0.32$. The prior says that we believe $\theta = 0.5$ with high probability. After observing the data we change our belief to a lower value, but because of the prior its still higher than the maximum likelihood. As the number n of balls drawn is increased the two estimates will, however, converge to the same value.

Question 3. We observe data $(x^{(i)}, y^{(i)})$, $i = 1, 2, \ldots, n$ from n people, where $x^{(i)}$ is the persons height and $y^{(i)}$ is the persons weight.

1. Explain how to construct a linear regression model for this data.
2. Suppose we suspect that the weight of a person is not linearly related to their height but rather is related to the square root of their height. Explain how to modify the linear regression model to accommodate this.

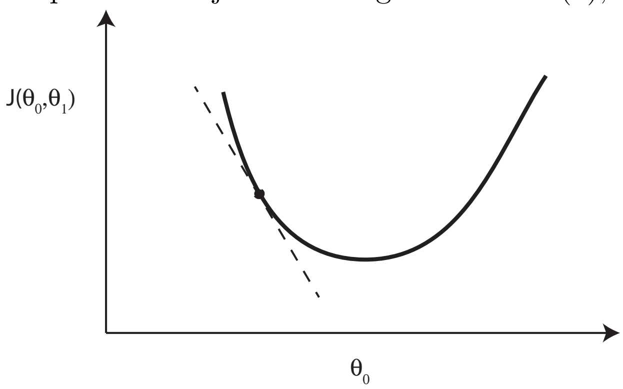
Solution

1. In a linear regression model we predict that the persons weight y given their height x is $h_\theta(x) = \theta x$, where θ is an unknown parameter (a single value since there is a single input x). To estimate the parameter we use the value which minimises the cost function $J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$.

2. We extend the input to be vector $x = [height, \quad \sqrt{height}]^T$. The prediction is now $h_\theta(x) = \theta^T x$ and we select the parameter vector which minimises $J(\theta)$ (using the new $h_\theta(x)$).

Question 4. Explain the principle of the gradient descent algorithm. Accompany your explanation with a diagram and pseudo-code.

Solution The task is to find the parameter vector θ which minimises the function $J(\theta)$. The basic idea is to iteratively update θ such that each update makes $J(\theta)$ smaller. One way to generate an update that does this is to use the gradient of $J(\theta)$. The gradient gives the slope of a line just touching the curve $J(\theta)$, e.g.



and so moving down this slope causes $J(\theta)$ to decrease. The resulting algorithm is:

- Start with some θ
- Repeat{
  for j=0 to n $\{tempj := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)\}$
  for j=0 to n $\{\theta_j := tempj\}$
  }
  where n is the number of elements in vector θ and $\alpha > 0$ is the learning rate. If α is selected too large then the algorithm may not converge, and if too small then convergence be slow.