Linear Regression

- 线性回归概率解释
  - Probabilistic interpretation
  - ❶ 做出假设 $y^{(i)} = \vartheta^T x^{(i)} + \varepsilon^{(i)}$
  - 假设误差项 $\varepsilon^{(i)} \sim N(0, \sigma^2)$ $\varepsilon^{(i)} \sim N(0, \sigma^2)$
  - ❷ 则可以得到 $p(\varepsilon^{(i)}) = \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(\varepsilon^{(i)})^2}{2\sigma^2})$
  - ❸ 将 $y^{(i)} = \vartheta^T x^{(i)} + \varepsilon^{(i)}$ $y^{(i)} = \vartheta^T x^{(i)} + \varepsilon^{(i)}$
  - $p(y^{(i)}|x^{(i)};\vartheta) = \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(y^{(i)}-\vartheta^T x^{(i)})^2}{2\sigma^2}) \sim N(\vartheta^T x^{(i)}, \sigma^2)$
    - 该函数代表以theta为参数，在x情况下，出现y的概率
  - ❹ 对m个样本的概率函数取似然函数（likelihood function）得到
    - $L(\theta) = L(\theta;X,\vec{y}) = p(\vec{y}|X;\vartheta) = \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(y^{(i)}-\vartheta^T x^{(i)})^2}{2\sigma^2})$
  - ❺ we should choose theta to maximize L(theta)　(maximum likelihood)
  - 取对数似然函数
    - $$\begin{aligned} \ell(\theta) &= logL(\theta) \\ &= log\prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(y^{(i)}-\vartheta^T x^{(i)})^2}{2\sigma^2}) \\ &= \sum_{i=1}^{m} log\frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{(y^{(i)}-\vartheta^T x^{(i)})^2}{2\sigma^2}) \\ &= mlog\frac{1}{\sqrt{2\pi\sigma}} - \frac{1}{\sigma^2}\frac{1}{2}\sum_{i=1}^{m}(y^{(i)}-\vartheta^T x^{(i)})^2 \end{aligned}$$
  - 最大化函数也就是最小化 $\frac{1}{2}\sum_{i=1}^{m}(y^{(i)}-\vartheta^T x^{(i)})^2$
    - 这个式子就是我们之前定义的损失函数J(theta)
- 局部加权线性回归
  - Locally weighted linear regression
  - LOESS
  - 它是非参数学习算法
  - 参数与非参数学习算法
    - Parametric learning algorithm
      - 通过固定的参数来拟合数据
    - no-parametric learning algorithm
      - 不对目标函数的形式作出强烈假设，参数的数量随m的增长而增长
  - 线性回归的预测过程
    - $Fit\ \vartheta\ to\ minimize\ \Sigma_i(y^i - \vartheta^T x^i)^2$
    - $Output\ \vartheta^T x$
  - 对比
  - 局部加权线性回归的预测过程
    - $Fit\ \vartheta\ to\ minimize\ \Sigma_i w^i(y^i - \vartheta^T x^i)^2$
    - $Output\ \vartheta^T x$
    - w是权值，可定义为 $w^{(i)} = exp(-\frac{(x^{(i)}-x)^2}{2})$
      - $If\ |x^{(i)}-x|\ is\ small,\ then\ w^{(i)} \approx 1$
      - $If\ |x^{(i)}-x|\ is\ large,\ then\ w^{(i)} \approx 0$
      - 这样就实现了局部加权，即预测某点时，离这个点越近的x，贡献越大，离这个点越远的x，贡献越小
      - $w^{(i)} = exp(-\frac{(x^{(i)}-x)^2}{2\tau^2})$ τ:控制了权值随距离下降的速率