COMP 562 – Lecture 3

Finding μ^{MLE} of Gaussian Distribution

We left as an exercise a problem to come up with a maximum likelihood estimate for parameter μ of a Gaussian distribution

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

So we will do that now

Likelihood function is

$$\mathcal{L}(\mu,\sigma^2|\mathbf{x}) = \prod_{i=1}^{N} p(x_i|\mu,\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}$$

Log-likelihood function is

$$\log \mathcal{L}(\mu,\sigma^2|\mathbf{x}) = \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} = \sum_{i=1}^{N} \left[-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x_i-\mu)^2 \right]$$

Finding μ^{MLE} of Gaussian Distribution

Our recipe is:

1. Take the function you want to maximize:

$$f(\mu) = \sum_{i=1}^{N} \left[-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x_i-\mu)^2 \right]$$

1. Compute its first derivative: $\frac{\partial}{\partial\mu} f(\mu)$
2. Equate that derivative to zero and solve: $\frac{\partial}{\partial\mu} f(\mu) = 0$

The first derivative is

$$\frac{\partial}{\partial \mu} f(\mu) = \sum_{i=1}^{N} \left[\frac{1}{\sigma^2} (x_i - \mu) \right]$$

We equate it to zero and solve

$$\sum_{i=1}^{N} \left[\frac{1}{\sigma^2} (x_i - \mu) \right] = 0$$

$$\frac{\sum_{i=1}^{N} x_i}{N} = \mu$$

Finding ${\sigma^2}^{MLE}$ of Gaussian Distribution

Our recipe is:

1. Take the function you want to maximize:

$$f(\sigma^2) = \sum_{i=1}^{N} \left[-\frac{1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} (x_i - \mu)^2 \right]$$

1. Compute its first derivative: $\frac{\partial}{\partial \sigma^2} f(\sigma^2)$
2. Equate that derivative to zero and solve: $\frac{\partial}{\partial \sigma^2} f(\sigma^2) = 0$

$$f(\sigma^2) = \sum_{i=1}^{N} \left[-\frac{1}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(x_i - \mu)^2 \right] = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left[(x_i - \mu)^2 \right]$$

The first derivative is

$$\frac{\partial}{\partial \sigma^2} f(\sigma^2) = -\frac{N}{2\sigma^2} - \left(\frac{1}{2} \sum_{i=1}^{N} \left[(x_i - \mu)^2 \right] \right) \frac{\partial}{\partial \sigma^2} \left(\frac{1}{\sigma^2} \right)$$

$$= -\frac{N}{2\sigma^2} - \left(\frac{1}{2} \sum_{i=1}^{N} \left[(x_i - \mu)^2 \right] \right) \left(-\frac{1}{(\sigma^2)^2} \right) = \frac{1}{2\sigma^2} \left(\frac{1}{\sigma^2} \sum_{i=1}^{N} \left[(x_i - \mu)^2 \right] - N \right)$$

Which, if we rule out $\sigma^2 = 0$, is equal to zero only if

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \left[(x_i - \mu)^2 \right]$$

Please Verify both μ^{MLE} and $\sigma^{2^{MLE}}$ using seconed derivative test

Linear Regression

Formally, we would write

$$y = \beta_0 + \sum_j x_j \beta_j + \epsilon$$
$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

or more compactly

$$y|\mathbf{x} \sim \mathcal{N}\left(\beta_0 + \sum_j x_j \beta_j, \sigma^2 \right)$$

Notice that the function is linear in the parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, not necessarily in terms of the covariates

Linear Regression

Probability of target variable y

$$p(y|\mathbf{x}, \beta_0, \beta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left(y_i - \underbrace{(\beta_0 + \sum_j x_j \beta_j)}_{\text{mean of the Gaussian}} \right)^2 \right\}$$

In the case of the 6th grader's height, we made **the same** prediction for any other 6th grader (58.5 inches)

In our COMP 562 grade example, we compute a potentially different mean for every student

$$\beta_0 + \beta_{\text{COMP410}} * \text{COMP410} + \beta_{\text{MATH233}} * \text{MATH233} + \beta_{\text{STOR435}} * \text{STOR435} + \beta_{\text{beers}} * \text{beers}$$

Linear Regression -- Likelihood

We start by writing out a likelihood for linear regression is

$$\mathcal{L}(\beta_0, \beta, \sigma^2|\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{N} p(y|\mathbf{x}, \beta_0, \beta, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left(y_i - (\beta_0 + \sum_j x_j \beta_j) \right)^2 \right\}$$

Log-likelihood for linear regression is

$$\log \mathcal{L}(\beta_0, \beta, \sigma^2|\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \left[-\frac{1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \left(y_i - (\beta_0 + \sum_j x_j \beta_j) \right)^2 \right]$$

$$= -\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(y_i - (\beta_0 + \sum_j x_{i,j} \beta_j) \right)^2 = -\frac{N}{2} \log(2\pi\sigma^2) - \frac{RSS}{2\sigma^2}$$

We will refer to expression $y_i - (\beta_0 + \sum_j x_j\beta_j)$ as **residual**, and hence **RSS** stands for **residual sum of squares** or **sum of squared errors** and is defined by

$$RSS = \sum_{i=1}^{N} \left(y_i - (\beta_0 + \sum_j x_{i,j}\beta_j) \right)^2$$

And RSS/N is called the **mean squared error** or **MSE**

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - (\beta_0 + \sum_j x_{i,j}\beta_j) \right)^2$$

Hence, maximizing log-likelihood is equivalent to minimizing RSS or MSE

Another way to see this is to consider a very simplified version of Taylor's theorem

Theorem. Given a function $f(\cdot)$ which is smooth at x

$$f(x + d) = f(x) + f'(x)d + O(d^2)$$

In words, close to x function $f(\cdot)$ is very close to being a linear function of d

$$f(x + d) = f(x) + f'(x)d$$

Slope of the best linear approximation is $f'(x)$, i.e., $f'(x)$ tells us in which direction function grows

- Gradient Ascent\Descent: Choose initial $\theta^{(0)} \in \mathbb{R}^n$, repeat:
  $$\theta^{(k)} = \theta^{(k-1)} \pm t_k . \nabla f(\theta^{(k-1)}), k = 1, 2, 3, \ldots$$
  Where t_k is the step size (learning rate) at step k
- Stop at some point using a stopping criteria (depend on the problem we are solving), for example:
  - Maximum number of iterations reached
  - $|f(\theta^{(k)}) - f(\theta^{(k-1)})| < \epsilon$

1. Use Line search Strategy

   - At each iteration, do the best you can along the direction of the gradient,
     $$t = \underset{s \geq 0}{\operatorname{argmax}} f(\theta + s . \nabla f(\theta))$$
   - Usually, it is not possible to do this minimization exactly, and approximation methods are used
   - Backtracking Line Search:
     - Choose an initial learning rate ($t_k = t_{init}$), and update your parameters
       $\theta^{(k)} = \theta^{(k-1)} \pm t_k . \nabla f(\theta^{(k-1)})$
     - Reduce learning rate $t_k = \alpha . t_{init}$, where $0 < \alpha < 1$
     - Repeat by reducing α till you see an improvmnet in $f(\theta^{(k)})$

Linear Regression -- Likelihood

We start by writing out a likelihood for linear regression is

$$\mathcal{L}(\beta_0, \beta, \sigma^2 | \mathbf{x}, \mathbf{y}) = \prod_{i=1}^{N} p(y | \mathbf{x}, \beta_0, \beta, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2\sigma^2} \left(y_i - (\beta_0 + \sum_j x_j \beta_j) \right)^2 \right\}$$

Log-likelihood for linear regression is

$$\log \mathcal{L}(\beta_0, \beta, \sigma^2 | \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} \left[-\frac{1}{2} \log 2\pi\sigma^2 - \frac{1}{2\sigma^2} \left(y_i - (\beta_0 + \sum_j x_j \beta_j) \right)^2 \right].$$

Linear Regression -- Gradient of Log-Likelihood

Partial derivatives

$$\frac{\partial}{\partial \beta_0} \log \mathcal{L}(\beta_0, \beta, \sigma^2 | \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} -\frac{1}{\sigma^2} \left(y_i - (\beta_0 + \sum_j x_j \beta_j) \right) (-1)$$

$$\frac{\partial}{\partial \beta_k} \log \mathcal{L}(\beta_0, \beta, \sigma^2 | \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{N} -\frac{1}{\sigma^2} \left(y_i - (\beta_0 + \sum_j x_j \beta_j) \right) (-x_k) \qquad , k \in \{1, \dots, p\}$$

Hence gradient (with respect to βs)

$$\nabla \log \mathcal{L}(\beta_0, \beta, \sigma^2 | \mathbf{x}, \mathbf{y}) = \begin{bmatrix} \sum_{i=1}^{N} -\frac{1}{\sigma^2} \left(y_i - (\beta_0 + \sum_j x_j \beta_j) \right) (-1) \\ \sum_{i=1}^{N} -\frac{1}{\sigma^2} \left(y_i - (\beta_1 + \sum_j x_j \beta_j) \right) (-x_1) \\ \vdots \\ \sum_{i=1}^{N} -\frac{1}{\sigma^2} \left(y_i - (\beta_0 + \sum_j x_j \beta_j) \right) (-x_p) \end{bmatrix}$$