Machine Learning 1 - Exercise 2

Fränz Beckius (374057)
Ivan David Aranzales Acero (399364)
Janek Tichy (584200)
Jeremias Eichelbaum (358685)
Johannes Krause (395469)

October 29, 2018

1 Maximum-Likelihood Estimation

1.a

$x, y \in R_+^2$ and p(x),p(y) are exponential distribution density function

$$p(x,y) = \lambda\eta \cdot e^{-\lambda x - \eta y} = \lambda e^{-\lambda x} \cdot \eta e^{-\eta y} = p(x) \cdot p(y)$$

$$\ln(p(x,y)) = \ln(\lambda\eta) - \lambda x - \eta y = -\lambda x + \ln(\lambda) - \eta y + \ln(\eta) = \ln(p(x) \cdot p(y))$$

$p(x,y) = p(x)p(y)$, so x and y are independent.

1.b

$$p(x, y \mid \lambda, \eta) = \lambda\eta \cdot e^{-\lambda x - \eta y}$$

with unknown λ and $\eta = \frac{1}{\lambda}$ we solve
given $\mathcal{D}$ and $\eta \in R_+$

$$min_\lambda = l(\lambda, \mathcal{D})$$

where

$$l(\lambda, \mathcal{D}) = -\sum_{n=1}^{N} \ln(p(x, y \mid \lambda))$$

$$= argmin_\lambda \left(-\sum_{n=1}^{N} \ln(\lambda\eta) + (-\lambda x_n - n y_n) \right)$$

$$= argmin_\lambda \left(-\sum_{n=1}^{N} \ln(\lambda\eta) - \lambda x_n + const.) \right)$$

$$0 = \nabla l(\lambda, \mathcal{D}) = -N \ln(\lambda) + N\lambda\bar{x}$$

$$\nabla l = -\frac{N}{\lambda} + Nx$$

$$Nx = \frac{N}{\lambda} \Rightarrow \bar{x} = \frac{1}{\lambda}$$

1.c

$$argmin_\lambda = \sum_{n=1}^{N} \lambda x_1 \cdot \frac{y_n}{\lambda}$$

$$\nabla\lambda = \sum_{n=1}^{N} \left(x_n + \frac{y_n}{\lambda^2}\right) = N\bar{x} \cdot \frac{N\bar{y}}{\lambda^2}$$

$$\Rightarrow \lambda = \sqrt{\frac{\bar{y}}{\bar{x}}}$$

1.d

$$argmin_\lambda l(\lambda, \mathcal{D}) = -\sum_{n=1}^{N} \ln(p(x_n, y_n \mid \lambda))$$

$$= -\sum_{n=1}^{N} (\ln(\lambda(1-\lambda) + (-\lambda x_n - (1-\lambda)y_n))$$

$$= -\sum_{n=1}^{N} (\ln(\lambda(1-\lambda) - \lambda x_n - (1-\lambda)y_n)$$

$$= -\sum_{n=1}^{N} (\ln(\lambda) + \ln(1-\lambda) - \lambda x_n - y_n + \lambda y_n)$$

$$0 = \nabla l = -\sum_{n=1}^{N} \left(\frac{1}{\lambda} + \frac{1}{1-\lambda} + y_n - x_n\right)$$

$$-\sum_{n=1}^{N} \left(\frac{1}{\lambda} + \frac{1}{1-\lambda} = x_n - y_n\right)$$

$$\frac{N}{\lambda} + \frac{N}{1-\lambda} = N\bar{x} - N\bar{y}$$

$$\lambda - \lambda^2 = \frac{1}{\bar{x} - \bar{y}}$$

$$\lambda = \frac{1}{2} \pm \frac{\sqrt{(-4 + \bar{x} - \bar{y})(\bar{x} - \bar{y})}}{2(\bar{x} - \bar{y})}$$

2 Maximum Likelihood vs. Bayes

2.a

$P(x_n \mid D)$ can be rewritten as $\theta^{x_n}(1-\theta)^{1-x_n}$ where $x_n \in 0{,}1$

$$argmin_\theta = \nabla l(\theta, \mathcal{D}) = \widehat{x}\theta^{\widehat{x}-1} \tag{1}$$

Where $\widehat{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$

$$0 = \nabla l = log(\widehat{x}) + log(\theta^{\widehat{x}-1}) \tag{2}$$
$$= log(\widehat{x}) + (\widehat{x}-1)log(\theta) \tag{3}$$
$$= exp(\frac{-log(\widehat{x})}{\widehat{x}-1}) = \theta \tag{4}$$

2.b

For $x_n = \frac{5}{7}$ we can compute $\theta = exp(\frac{-log(\frac{5}{7})}{-\frac{2}{5}}) = 0.5999$

Since the two events are independent we can compute the probability for $x_8, x_9 = head$ like this:

$$p(x_8 = head, x_9 = head|\theta) = p(x_8 = head|\theta)p(x_9 = head|\theta) = 0.5999^2 = 0.3598 \tag{5}$$

Since $p(\theta) = 1$ for $\theta \in [0, 1]$, we can reformulate:

$$\int P(x_8 = head, x_9 = head \mid theta)p(\theta \mid D)d\theta \tag{6}$$
$$= \int P(x_8 = head \mid \theta)P(x_9 = head \mid theta)p(\theta \mid D)d\theta \tag{7}$$
$$= \int P(x_8 = head \mid \theta)P(x_9 = head \mid theta)p(D \mid \theta)d\theta \tag{8}$$

Now we compute $p(\theta \mid D)$

$$p(\theta \mid D) = \prod_{n=1}^{N} \theta^{x_n}(1-\theta)^{1-x_n} \tag{9}$$
$$= \sum_{n=1}^{N} log(\theta^{x_n}(1-\theta)^{1-x_n}) \tag{10}$$
$$= \sum_{n=1}^{N} (1-x_n)log(1-\theta) + x_n log(\theta) \tag{11}$$
$$= (1-\widehat{x})log(1-\theta) + \widehat{x}log(\theta) \tag{12}$$

2.c

3 Convergence of Bayes Parameter Estimation

3.a

$$\frac{1}{\sigma_n^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \iff \sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

$$\lim_{n\to\infty} \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2} = \frac{\sigma^2}{n}$$

If we receive a large quantity of new data, this will diminish the importance of the variance of our prior guess. However, if our uncertainty of the mean based on the prior data is non existent (variance = 0), no additional data will further reduce the posterior variance.
$\implies \sigma_n^2 \leq min(\frac{\sigma^2}{n}, \sigma_0^2)$

3.b

$$\frac{\mu}{\sigma_n^2} = \frac{n}{\sigma^2}\bar{x}_n + \frac{\mu_0}{\sigma_0^2}$$

$$\iff \mu_n = \frac{n\sigma_n^2}{\sigma^2}\bar{x}_n + \frac{\sigma_n^2}{\sigma_0^2}\mu_0$$

Since:

$$\sigma_n^2 = \frac{\sigma_0^2\sigma^2}{n\sigma_0^2 + \sigma^2}$$

$$\implies \mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\bar{x}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

Now we replace:

$$n\sigma_0^2 = a, \sigma^2 = b$$

$$\implies \mu_n = \frac{a}{a+b}\bar{x}_n + \frac{b}{a+b}\mu_0$$

As a result of the linear relation, if:

$$\bar{x}_n < \mu_0 \implies \bar{x}_n \leq \mu_n \leq \mu_0$$

$$\bar{x}_n \geq \mu_0 \implies \mu_0 \leq \mu_n \leq \bar{x}_n$$

Therefore:

$$min(\bar{x}_n, \mu_0) \leq \mu_n \leq max(\bar{x}_n, \mu_0)$$