Data Science Supervision 1

5. PDF of a normal distribution is $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = f(x)$

$\therefore \log(f(x)) = \text{constant} - \log(\sigma) - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$

$\log(\text{likelihood}) = \left(\sum_{i=1}^{m} \log(f_A(x_i))\right) + \left(\sum_{i=1}^{n} \log(f_B(y_i))\right)$

assuming independence

$= \left(\sum_{i=1}^{m} -\log(\sigma) - \frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right) + \left(\sum_{i=1}^{n} -\log(\sigma) - \frac{1}{2}\left(\frac{y_i-\mu-\delta}{\sigma}\right)^2\right)$

+ constant

~~$= (-m-n)\log\sigma - \frac{1}{2\sigma^2}\left(\left(\sum_{i=1}^{m} x_i^2 + \mu^2 - 2\mu x_i\right) + \sum_{i=1}^{n}\right)$~~

$= \text{constant} - (m+n)\log\sigma - \frac{1}{2\sigma^2}\left(\left(\sum_{i=1}^{m}(x_i-\mu)^2\right) + \left(\sum_{i=1}^{n}(y_i-\mu-\delta)^2\right)\right)$

$\frac{\partial \log(\text{likelihood})}{\partial \delta} = -\frac{1}{2\sigma^2}\sum_{i=1}^{n} -2(y_i-\mu-\delta)$

$= \frac{1}{\sigma^2}(n\bar{y} - n\mu - n\delta)$

$\vdots = 0$

$\therefore \bar{y} - \mu - \delta = 0$

$\therefore \mu + \delta = \bar{y}$

$\frac{\partial \log(\text{likelihood})}{\partial \mu} = -\frac{1}{2\sigma^2}\left(\left(\sum_{i=1}^{m} -2(x_i-\mu)\right) + \left(\sum_{i=1}^{n} -2(y_i-\mu-\delta)\right)\right)$

$= \frac{1}{\sigma^2}(m\bar{x} - m\mu + n\bar{y} - n\mu - n\delta)$

$= \frac{1}{\sigma^2}(m\bar{x} - m\mu + 0)$

$\vdots = 0$

$\therefore \bar{x} = \mu \Rightarrow \delta = \bar{y} - \bar{x}$

$$\frac{\partial \log(\text{likelihood})}{\partial \sigma} = -\frac{m+n}{\sigma} + \frac{1}{4\sigma^3}\left(\left(\sum_{i=1}^{m}(x_i-\mu)^2\right) + \left(\sum_{i=1}^{n}(y_i-\mu-\delta)^2\right)\right)$$

$$\stackrel{!}{=} 0$$

$$\therefore 4\sigma^2(m+n) - \left(\left(\sum_{i=1}^{m}(x_i^2+\mu^2-2\mu x_i)\right) + \left(\sum_{i=1}^{n}(y_i^2+(\mu+\delta)^2-2(\mu+\delta)y_i)\right)\right) = 0$$

$$\therefore 4\sigma^2(m+n) - \left(m\overline{x^2} + m\bar{x}^2 - 2m\bar{x}^2 + n\overline{y^2} + n\bar{y}^2 - 2n\bar{y}^2\right) = 0$$

$$\therefore 4\sigma^2 = \frac{m(\overline{x^2}-\bar{x}^2) + n(\overline{y^2}-\bar{y}^2)}{n+m}$$

$$\therefore \sigma = \frac{1}{2}\sqrt{\frac{m(\overline{x^2}-\bar{x}^2) + n(\overline{y^2}-\bar{y}^2)}{n+m}}$$

7. Using feature vectors:

$$\underbrace{\overrightarrow{x1}_{x<\text{inflection}},\ \overrightarrow{x1}_{x\geq\text{inflection}}}_{\text{element-wise multiplication}},\ \overrightarrow{1}_{x<\text{inflection}},\ \overrightarrow{1}_{x\geq\text{inflection}}$$

The linear model can be written as

$$y = m_1\overrightarrow{x1}_{x<\text{inflection}} + c_1\overrightarrow{1}_{x<\text{inflection}} + m_2\overrightarrow{x1}_{x\geq\text{inflection}} + c_2\overrightarrow{1}_{x\geq\text{inflection}}$$

Constrained by the equation $m_1 \cdot \text{inflection} + c_1 = m_2 \cdot \text{inflection} + c_2$

$$\therefore c_2 = (m_1-m_2)\cdot\text{inflection} + c_1$$

Let i = inflection.

$$\therefore y = m_1\overrightarrow{x1}_{x<i} + c_1\overrightarrow{1}_{x<i} + m_2\overrightarrow{x1}_{x\geq i} + m_1\cdot i\overrightarrow{1}_{x\geq i} - m_2 i\overrightarrow{1}_{x\geq i} + c_1\overrightarrow{1}_{x\geq i}$$

$$= m_1\left(\overrightarrow{x1}_{x<i} + i\overrightarrow{1}_{x\geq i}\right) + m_2\left(\overrightarrow{x1}_{x\geq i} - i\overrightarrow{1}_{x\geq i}\right) + c_1\left(\overrightarrow{1}_{x<i} + \overrightarrow{1}_{x\geq i}\right)$$

With new feature vectors

$$A = x\vec{1}_{x<i} + i\vec{1}_{x\geq i},$$
$$B = x\vec{1}_{x\geq i} + i\vec{1}_{x<i}$$
$$C = \vec{1}$$

The model can be written as

$$y = m_1 A + m_2 B + c_1 C$$

where $c_2 = c_1 + (m_1 - m_2)i$

8. Let $\vec{\delta}_k = \vec{1}_{u=k}$

$$temp \sim \vec{\alpha} + \beta_1 \sin(2\pi\vec{t}) + \beta_2 \cos(2\pi\vec{t}) + \varepsilon_1 \vec{\delta}_{decade\_1930s} + \varepsilon_2 \vec{\delta}_{decade\_1990s} + \varepsilon_3 \vec{\delta}_{decade\_2000s} + \varepsilon_4 \vec{\delta}_{decade\_2010s} + \varepsilon_5 \vec{\delta}_{decade\_2020s}$$

Feature vectors are $\{\vec{a}, \sin(2\pi\vec{t}), \cos(2\pi\vec{t})\} \cup \{\vec{\delta}_k \mid k \in \{decade\_1980s \dots decade\_2020s\}\}$

It can be solved using the attached code.

```
(a, b1, b2, e1, e2, e3, e4, e5) = linear_model.LinearRegression( fit_intercept=False).fit(
    numpy.column.stack(
        numpy.ones(len(temp)),
        numpy.sin(2 * numpy.pi * t),
        numpy.cos(2 * numpy.pi * t),
        numpy.where(u="decase_1980s", 1, 0),
        numpy.where(u="decase_1990s", 1, 0),
        numpy.where(u="decase_2000s", 1, 0),
        numpy.where(u="decase_2010s", 1, 0),
        numpy.where(u="decase_2020s", 1, 0)
    ),
    temp
)
```

10. Let $\vec{\gamma}_k = \vec{1}_{gender=k}$,

$\vec{\delta}_k = \vec{1}_{eth=k}$

$$\vec{1}_{outcome=\text{"find"}} = \varepsilon_1 \vec{\delta}_{Asian} + \varepsilon_2 \vec{\delta}_{Black} + \varepsilon_3 \vec{\delta}_{Mixed} + \varepsilon_4 \vec{\delta}_{other} + \varepsilon_5 \vec{\delta}_{white} + \varepsilon_6 \vec{\gamma}_{female} + \varepsilon_7 \vec{\gamma}_{male}$$

the parameters are not identifiable because

$$\sum_{eth} \vec{\delta} = \sum_{gender} \vec{\gamma}$$

A linearly independent model would be:

$$\vec{1}_{outcome=\text{"find"}} = \varepsilon_1 \vec{\delta}_{Asian} + \varepsilon_2 \vec{\delta}_{Black} + \varepsilon_3 \vec{\delta}_{Mixed} + \varepsilon_4 \vec{\delta}_{other} + \varepsilon_5 \vec{\delta}_{white} + \varepsilon_6 \vec{\gamma}_{female} + \varepsilon_7 \left(\sum_{eth} \vec{\delta} - \vec{\gamma}_{female}\right)$$

$$= (\varepsilon_1 + \varepsilon_7)\vec{\delta}_{Asian} + (\varepsilon_2 + \varepsilon_7)\vec{\delta}_{Black} + (\varepsilon_3 + \varepsilon_7)\vec{\delta}_{Mixed} + (\varepsilon_4 + \varepsilon_7)\vec{\delta}_{other} + (\varepsilon_5 + \varepsilon_7)\delta_{white} + (\varepsilon_6 - \varepsilon_7)\gamma_{female}$$

where ε_1 = bias towards Asian people
ε_2 = bias towards Black people
ε_3 = bias towards Mixed race people
ε_4 = bias towards people of other ethnicities
ε_5 = bias towards white people
ε_6 = bias towards females
ε_7 = bias towards males