Maximum Likelihood Estimation (MLE)

Bayes Rule Application

- Bayes Rule:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum p(x|y)p(y)}$$

- Assume the test has 80% sensitivity, i.e., if a patient has cancer then the test will come back positive with probability 0.8
  - $p(x = 1|y = 1) = 0.8$
    - $x = 1$: mammogram is positive
    - $y = 1$: patient has cancer
  - It does not mean the patient is 80% likely to have cancer
- The test may generate false positive (false alarm)
  - $p(x = 1|y = 0) = 0.2$
- Assume the prior probability $p(y = 1) = 0.004$

Bayes Rule Application

- Bayes Rule:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum p(x|y)p(y)}$$

- Given:
  - $p(x = 1|y = 1) = 0.8$
  - $p(x = 1|y = 0) = 0.2$
  - $p(y = 1) = 0.004 \rightarrow p(y = 0) = 0.996$

- From the mammogram, what is the probability that a patient has cancer ?
  - $p(y = 1|x = 1) = \frac{p(x=1|y=1)p(y=1)}{p(x=1|y=1)p(y=1)+p(x=1|y=0)p(y=0)} = 0.016$
  - If the test is positive, there is a chance of ~1.6% that the patient has cancer

Bayes Rule in Machine Learning

- Bayes Rule:

$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(\mathbf{x}|y = c, \boldsymbol{\theta})p(y = c|\boldsymbol{\theta})}{\sum_{c'} p(\mathbf{x}|y = c', \boldsymbol{\theta})p(y = c'|\boldsymbol{\theta})}$$

$\mathbf{x}$: feature vector; $\boldsymbol{\theta}$: parameter vector; y : class label

- In words,

$$posterior = \frac{likelihood \text{ x } prior}{evidence}$$

  - evidence is just a normalizing factor
  - $posterior = \ likehood$ x $prior$

- Generative Classifier → learns from the conditional probability and the prior probability
- Discriminative Classifier → learns directly from $p(y = c|\mathbf{x})$

Likelihood Function

- $p(D|\boldsymbol{\theta})$: likelihood function of $\boldsymbol{\theta}$ with respect to a set of examples in data set D

- MLE chooses the value of $\boldsymbol{\theta}$ that maximizes $p(D|\boldsymbol{\theta})$
  - $\boldsymbol{\theta} = argmax_{\theta}\, p(D|\boldsymbol{\theta})$

- D contains m examples drawn independently
  - Hence, $p(D|\boldsymbol{\theta}) = \prod_{i=1}^{m} p(\mathbf{x}_i|\boldsymbol{\theta})$

Log Likelihood Function

- Easier to work with log likelihood

$$\ell(\boldsymbol{\theta}) = \log\, p(D|\boldsymbol{\theta}) = \sum_{i=1}^{m} \log(p(\mathbf{x}_i|\boldsymbol{\theta}))$$

- Since the log function is monotonically increasing, $\boldsymbol{\theta}$ that maximizes the log likelihood also maximizes the likelihood function

$$\boldsymbol{\theta} = \mathrm{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$$

$$\nabla \ell(\boldsymbol{\theta}) = \nabla \log\, p(D|\boldsymbol{\theta}) = \sum_{i=1}^{m} \nabla \log\big(p(\mathbf{x}_i|\boldsymbol{\theta})\big) = 0$$

- The solution to this equation is $\boldsymbol{\theta}_{MLE}$, which maximizes the likelihood function

Univariate & Multivariate Gaussian

- For univariate Gaussian: $x_i \sim \mathcal{N}(\mu, \sigma^2)$

$$\mu_{MLE} = \frac{1}{m}\sum_{i=1}^{m} x_i = \bar{x}$$

$$\sigma^2{}_{MLE} = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^2$$

- For multivariate Gaussian: $\mathbf{x}^{(i)} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

$$\boldsymbol{\mu}_{MLE} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}^{(i)} = \bar{\mathbf{x}}$$

$$\Sigma_{MLE} = \frac{1}{m}\sum_{i=1}^{m}\left(\mathbf{x}^{(i)} - \bar{\mathbf{x}}\right)\left(\mathbf{x}^{(i)} - \bar{\mathbf{x}}\right)^T = \frac{1}{m}(X - M)^T(X - M)$$

  - M is the mean matrix

- MLE of the mean vector is the sample mean
- MLE of the covariance matrix is the arithmetic average of m matrices $\left(\mathbf{x}^{(i)} - \bar{\mathbf{x}}\right)\left(\mathbf{x}^{(i)} - \bar{\mathbf{x}}\right)^T$