02/10/2016

Gaussian Discriminant Analysis

- Fundamental assumption: each class comes from a normal distribution (Gaussian).

$$X \sim \mathcal{N}(\mu, \sigma^2) : P(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{|x-\mu|^2}{2\sigma^2}}$$

- For each class c, suppose we estimate mean μ_c, variance σ_c^2, and prior $\pi_c = P(Y = c)$.
- Given x, Bayes' rule $r^*(x)$ return class C that maximizes $P(X = x|Y = c)\pi_c$.
- $\ln z$ is monotonically increasing for $z > 0$, so its equivalent to maximize,

$$\begin{aligned} Q_c(x) &= \ln(\sqrt{2\pi} P(X = c|Y = c)\pi_c) \\ &= -\frac{|x - \mu_c|^2}{2\sigma^2} - \ln \sigma_c + \ln \pi_c \end{aligned}$$

- $Q_c(x)$ is quadratic in x.

Quadratic Discriminant Analysis (QDA)

- Suppose only 2 classes c, d, Then,

$$r^*(x) = \begin{cases} c & \text{if } Q_c(x) - Q_d(x) > 0 \\ d & \text{otherwise} \end{cases}$$

- The $Q_c(x) - Q_d(x)$ prediction function is quadratic in x.
- Bayes decision boundary is $Q_c(x) - Q_d(x) = 0$.
- In 1D, Bayesian decision boundary may have 1 or 2 points.
- In d-D, Bayesian decision boundary is a quadric.
- To recover posterior probabilities in 2-class case, use Bayes:

$$P(Y = c|X) = \frac{P(X|Y = c)\pi_c}{P(X|Y = c)\pi_c + P(X|Y = d)\pi_d}$$

- Recall $e^{Q_c(x)} = \sqrt{2\pi} P(x)\pi_c$.

$$\begin{aligned} P(Y = c|X = x) &= \frac{e^{Q_c(x)}}{e^{Q_c(x)} + e^{Q_d(x)}} \\ &= \frac{1}{1 + e^{Q_c(x) - Q_d(x)}} \\ &= s(Q_c(x) - Q_d(x)) \end{aligned}$$

- Where s(·) is the logistic function aka sigmoid function,

$$s(\gamma) = \frac{1}{1 + e^{-\gamma}}$$

- Monotonically increasing.
- $s(0) = \frac{1}{2}$.
- $s(\infty) \to 1$.
- $s(-\infty) \to -1$.
- always $\in [0,1] \to$ probabilities.

Linear Discriminant Analysis (LDA)

- Fundamental assumption: all the Gaussians have the same variance σ only difference between classes is the mean μ_i.
- Then,

$$Q_c(x) - Q_d(x) = \frac{(\mu_c - \mu_d) \cdot x}{\sigma^2} - \frac{\mu_c^2 - \mu_d^2}{2\sigma^2} + \ln \pi_c + \ln \pi_d$$

- Now its a linear classifiers! Choose c that maximizes,

$$\frac{\mu_c \cdot x}{\sigma^2} - \frac{\mu_c^2}{2\sigma^2} + \ln \pi_c$$

- In 2-class case, decision boundary is $w \cdot x + \alpha = 0$.
- If $\pi_c = \pi_d = \frac{1}{2} \implies (\mu_c - \mu_d) \cdot x - \frac{(\mu_c - \mu_d)}{2} = 0$
- This is the centroid method!
- In 2-class case, Bayes posterior is $P(Y = c|X = x) = s(w \cdot x + \alpha)$

Maximum Likelihood Estimation of Parameters

(Ronald Fisher, circa 1912)

- Lets flip biased coins. Heads with probability p; tails with probability $1 - p$.
- 10 flips, 8 heads, 2 tails. What is the most likely value of p?
- Binomial Distribution: $X \sim B(n, p)$

$$P[X = x] = \binom{n}{x} p^x (1-p)^{n-x}$$

- Our example: n=10,

$$P[X = 10] = 45p^8(1-p)^2 = \mathcal{L}(x)$$

- Probability of 8 heads in 10 flips: written as a function $\mathcal{L}(p)$ of distribution parameter(s); this is the <u>likelihood function</u>

- Maximum likelihood estimation (MLE): A method of estimating parameters of a statistical model by picking the parameters that maximize the likelihood function.

$$\boxed{\text{Find } p \text{ that maximizes } \mathcal{L}(p)}$$

- Solve this example by setting derivative equal to 0:

$$\frac{d\mathcal{L}}{dp} = 360p^7(1-p)^2 - 90p^8(1-p) = 0$$
$$\implies 4(1-p) - p = 0 \implies p = 0.8$$

- The log likelihood $\mathcal{L}(\cdot)$ is the ln of the likelihood $\mathcal{L}(\cdot)$.

Likelihood of a Gaussian

- Given samples $x_1, x_2, \ldots, x_n$ find best-fit Gaussian.
- Likelihood of generating these samples is,

$$\mathcal{L}(\mu, \sigma; x_1, \ldots, x_n) = P(x_1)P(x_2)\ldots P(x_n)$$

- Log likelihood is,

$$l(\mu, \sigma) = \sum_{i=1}^{n} \ln P(x_i)$$

- Want to set $\nabla_\mu l = 0$, and $\frac{\partial l}{\partial \sigma} = 0$.
- Natural log of Gaussian distribution,

$$\ln P(x) = -\frac{|x-\mu|^2}{2\sigma^2} - \ln \sqrt{2\pi} - \ln \sigma$$

- taking the gradient,

$$\nabla_\mu l = \sum_i \frac{x_i - \mu}{\sigma^2} \implies \hat{\mu} = \frac{1}{n}\sum_i x_i$$
$$\frac{\partial l}{\partial \sigma} = \sum_i \frac{|x_i - \mu|^2 - \sigma^2}{\sigma^3} = 0 \implies \hat{\sigma^2} = \frac{1}{n}\sum_i |x_i - \mu|^2$$

- We don't know μ exactly, so substitute $\hat{\mu}$ in the last equation above.
- For QDA: estimate mean and variance of each class as above, and estimate the priors (for each class c):

$$\hat{\pi_c} = \frac{n_c}{\sum_d n_d} \leftarrow \text{ denominator is the sum of samples in all classes}$$

- For LDA: same mean and priors; one variance for all classes:

$$\hat{\sigma^2} = \frac{1}{n}\sum_c \sum_{i:y_i=c} |x_i - \mu_c|^2$$