University of California San Diego

ECE 271 Notes

Instructor: *Prof. Nuno Vasconcelos*

Organized by Ray Tsai

Bayes Decision Rule

$$
\begin{aligned}
g^*(x) &= \arg\min_{g(x)} \sum_i P_{Y|X}(i|x) L[g(x), i] \\
&= \arg\max_i P_{Y|X}(i|x) && \text{(for 0-1 loss function)} \\
&= \arg\max_i P_{X|Y}(x|i) P_Y(i) && \text{(for 0-1 loss function)} \\
&= \arg\max_i \log P_{X|Y}(x|i) + \log P_Y(i). && \text{(for 0-1 loss function)}
\end{aligned}
$$

For binary classification, the likelihood ratio form is: pick 0 if $\frac{P_{X|Y}(x|0)}{P_{X|Y}(x|1)} > T^* = \frac{P_Y(1)}{P_X(0)}$.

Associated Risk

$$R^* = \int P_X(x) \sum_{i \neq g^*(x)} P_{Y|X}(i|x) dx = \int P_{Y,X}(y \neq g^*(x), x) dx \quad \text{(For 0-1 loss function)}$$

Gaussian Classifier

For single variable, we assume $\sigma_i = \sigma$ and pick 0 if

$$x < \frac{\mu_1 + \mu_0}{2} + \frac{1}{\frac{\mu_1 - \mu_0}{\sigma^2}} \log \frac{P_Y(0)}{P_Y(1)}.$$

Generalizing it to multiple variables, we assume $\Sigma_i = \Sigma$, then the BDR becomes

$$i^*(x) = \arg\min_i [d(x, \mu_i) + \alpha_i],$$

where $d(x, y) = (x - y)^T \Sigma^{-1} (x - y)$ and $\alpha_i = \cancel{\log\left[(2\pi)^d |\Sigma|\right]} - 2 \log P_Y(i)$.
Alternatively,

$$i^*(x) = \arg\max_i g_i(x),$$

where $g_i(x) = w_i^T x + w_{i0}$, $w_i = \Sigma^{-1} \mu_i$, and $w_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P_Y(i)$.

Geometric Interpretation

Thus, the hyperplane between class 0 and 1 is

$$g_0(x) - g_1(x) = w^T x + b = 0,$$

where $w = \Sigma^{-1}(\mu_0 - \mu_1)$ and $b = -\frac{(\mu_0 + \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1)}{2} + \log \frac{P_Y(0)}{P_Y(1)}$.
It could also be rewritten as

$$w^T (x - x_0) = 0,$$

where $w = \Sigma^{-1}(\mu_0 - \mu_1)$ and $x_0 = \frac{\mu_0 + \mu_1}{2} - \frac{1}{(\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1)} \log \frac{P_Y(0)}{P_Y(1)} (\mu_0 - \mu_1)$

Gaussian Distribution Transformation

Let $x \sim N(\mu, \Sigma)$, and let $y = A^T x$, for some matrix A. Then, $y \sim N(A^T \mu, A^T \Sigma A)$. A special case of this is the whitening transform $A_w = \Phi \Lambda^{-1/2}$, where Φ is the matrix of orthonormal eigenvectors of Σ, and Λ is the diagonal matrix of eigenvalues of Σ.

Sigmoid

Suppose that $g_1(x) = 1 - g_0(x)$. Then, we can rewrite

$$g_0(x) = \frac{1}{1 + \frac{P_{X|Y}(x|1)P_Y(1)}{P_{X|Y}(x|0)P_Y(0)}} = \frac{1}{1 + \exp\{d_0(x, \mu_0) - d_1(x, \mu_1) + \alpha_0 - \alpha_1\}},$$

where $d(x, y) = (x - y)^T \Sigma^{-1}(x - y)$ and $\alpha_i = \log\left[(2\pi)^d |\Sigma_i|\right] - 2\log P_Y(i)$.

Maximum Likelihood Estimation

Solve for

$$\theta^* = \arg\max_{\Theta} P_{X;\Theta}(\mathcal{D}; \theta) = \arg\max_{\Theta} \log P_{X|\Theta}(\mathcal{D}; \theta).$$

Consider the Gaussian example:

Given a sample $\mathcal{D} = \{x_1, \ldots, x_n\}$ of independent points, where $P_X(x_i) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} e^{-\frac{1}{2}(x_i - \mu)^T \Sigma^{-1}(x_i - \mu)}$.

Then, the likelihood $L(x_1, \ldots, x_n|\mu, \sigma) = \prod_{i=1}^n P_X(x_i)$. We take the gradient of the natural log of L with respect to μ and get

$$\begin{aligned}\nabla_\mu(\log L) &= \nabla_\mu\left(-\frac{1}{2}\log[(2\pi)^d|\Sigma|] - \frac{1}{2}\sum_{i=1}^n (x_i - \mu)^T \Sigma^{-1}(x_i - \mu)\right) \\ &= \sum_{i=1}^n \Sigma^{-1}(x_i - \mu) = \sum_{i=1}^n x_i - \sum_{i=1}^n \mu = 0.\end{aligned}$$

Thus, we get $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$. By taking the Hessian, we get $\nabla_\mu^2(\log L) = -\sum_{i=1}^n \Sigma^{-1} = -n\Sigma^{-1}$. Since the covariance matrix Σ is positive definite, $-n\Sigma^{-1}$ is negative definite. Thus $\hat{\mu}$ is the maximum point.

In addition, the MLE of the covariance matrix is

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T.$$

Bias and Variance

$$\begin{aligned}Bias(\hat{\theta}) = E[\hat{\theta} - \theta], \quad Var(\hat{\theta}) &= E\left\{(\hat{\theta} - E[\hat{\theta}])^2\right\}, \\ MSE(\hat{\theta}) = E\left[(\hat{\theta} - \theta)^2\right] &= Var(\hat{\theta}) + Bias^2(\hat{\theta}).\end{aligned}$$

Least Squares

Consider an overdetermined system $\Phi\theta = z$, where we attempt to minimize $\|z - \Phi\theta\|$, the least square solution is

$$\theta^* = (\Phi^T\Phi)^{-1}\Phi^T z.$$

For a overdetermined system $W\Phi\theta = Wz$, where we attempt to minimize $(z - \Phi\theta)^T W^T W(z - \Phi\theta)$, the least square solution is

$$\theta^* = (\Phi^T W^T W \Phi)^{-1}\Phi^T W^T W z.$$

Bayesian Estimation

Pick i if

$$i^*(x) = \arg\max_i P_{X|Y,T}(x|i, \mathcal{D}_i)P_Y(i),$$

where the class conditional is the predictive distribution

$$P_{X|Y,T}(x|i, \mathcal{D}_i) = \int P_{X|Y,\Theta}(x|i, \theta)P_{\Theta|Y,T}(\theta|i, \mathcal{D}_i)\, d\theta = E_{\Theta|Y,T}[P_{X|i,\Theta}(x|\theta) \,|\, T = \mathcal{D}_i].$$

For the multivariate Gaussian case, suppose

$$P_{T|\mu}(\mathcal{D}|\mu) = \mathcal{G}(\mathcal{D}, \mu, \Sigma), \quad P_\mu(\mu) = \mathcal{G}(\mu, \mu_0, \Sigma_0),$$

for known Σ, μ_0, Σ_0. The posterior distribution is $P_{\mu|T}(\mu|\mathcal{D}) = \mathcal{G}(\mu, \mu_n, \Sigma_n)$, where

$$\begin{aligned}
\Sigma_n &= \Sigma_0 A^{-1}\frac{1}{n}\Sigma \Rightarrow \Sigma_n^{-1} = n\Sigma^{-1} + \Sigma_0^{-1},\\
\mu_n &= \Sigma_0 A^{-1}\mu_{ML} + \frac{1}{n}\Sigma A^{-1}\mu_0,\\
A &= \Sigma_0 + \frac{1}{n}\Sigma.
\end{aligned}$$

Then, the predictive distribution is

$$\begin{aligned}
P_{X|T}(x|\mathcal{D}) &= \int P_{X|\mu}(x|\mu)P_{\mu|T}(\mu|\mathcal{D})\, d\mu\\
&= \int \mathcal{G}(x, \mu, \Sigma)\mathcal{G}(\mu, \mu_n, \Sigma_n)\, d\mu\\
&= \int \mathcal{G}(x - \mu, 0, \Sigma)\mathcal{G}(\mu, \mu_n, \Sigma_n)\, d\mu\\
&= \mathcal{G}(x, 0, \Sigma) * \mathcal{G}(x, \mu_n, \Sigma_n) = \mathcal{G}(x, \mu_n, \Sigma + \Sigma_n).
\end{aligned}$$

Note that for non-informative prior, $\lim_{|\Sigma_0|\to\infty} \mu_n = \mu_{ML}$ and $\lim_{|\Sigma_0|\to\infty} \Sigma_n = \frac{1}{n}\Sigma = \Sigma_{ML}$, so

$$P_{X|T}(x|\mathcal{D}) = \mathcal{G}(x, \mu_n, \Sigma + \Sigma_n) = \mathcal{G}\left(x, \mu_{ML}, \left(1 + \frac{1}{n}\right)\Sigma\right).$$

MAP Estimation

$$\theta_{MAP} = \arg\max_\theta P_{\Theta|T}(\theta|\mathcal{D}) = \arg\max_\theta P_{T|\Theta}(\mathcal{D}|\theta)P_\Theta(\theta),$$

and this makes the predictive distribution equal to

$$P_{X|T}(x|\mathcal{D}) = P_{X|\Theta}(x|\theta_{MAP}) = \mathcal{G}(x, \mu_{ML}, \Sigma)$$

Note that for the MAP estimator approaches the ML estimator as the sample size increases, i.e. $\theta_{MAP} \to \theta_{ML}$ as $n \to \infty$.

Expectation-maximization

1. write down the likelihood of the complete data (can drop terms irrelevant to Z and Ψ)

$$P_{X,Z}(\mathcal{D}, z; \Psi) = \left(\prod_{i=1}^{n} P_{X|Z}(x_i|z;\Psi)\right) P_Z(z;\Psi).$$

2. **E-step**: write down the Q function

$$Q(\Psi;\Psi^{(n)}) = E_{Z|X;\Psi^{(n)}}[\log P_{X,Z}(\mathcal{D}, z; \Psi) \,|\, \mathcal{D}].$$

3. **M-step**: update Ψ, i.e.

$$\Psi^{(n+1)} = \arg\max_{\Psi} Q(\Psi;\Psi^{(n)}).$$

EM for Mixtures

Represent the class variable as $z = e_j = (\underbrace{0,\ldots,1_j,\ldots,0}_{C \text{ entries}})^T$. The complete data log likelihood is

$$\log P_{X,Z}(\mathcal{D}, \{z_1,\ldots,z_n\};\Psi) = \log \prod_{i=1}^{n}\prod_{j=1}^{C}[P_{X|Z}(x|e_j,\Psi)\pi_j]^{z_{ij}} = \sum_{i,j} z_{ij}\log[P_{X|Z}(x|e_j,\Psi)\pi_j].$$

Thus, in E-step,

$$Q(\Psi;\Psi^{(n)}) = \sum_{i,j} h_{ij}\log[P_{X|Z}(x|e_j,\Psi)\pi_j]$$

where $h_{ij} = E_{Z|X;\Psi^{(n)}}[z_{ij}|\mathcal{D}]$. Hence, we only have to compute

$$h_{ij} = E_{Z|X;\Psi^{(n)}}[z_{ij}|\mathcal{D}] = P_{Z|X}(z_{ij}=1|x_i;\Psi^{(n)}) = P_{Z|X}(e_j|x_i;\Psi^{(n)})$$

In M-step, we compute

$$\Psi^{(n+1)} = \arg\max_{\Psi} \sum_{i,j} h_{ij}\log[P_{X|Z}(x|e_j,\Psi)\pi_j].$$

For Gaussiam mixure, we may solve for h_{ij} first then take the Lagrangian $L = Q(\Psi;\Psi^{(n)}) + \lambda\left(\sum_{j=1}^{C}\pi_j - 1\right)$ to solve for the parameters. Here are the results:

$$h_{ij} = \frac{\mathcal{G}\left(x_i,\mu_j^{(n)},\sigma_j^{(n)}\right)\pi_j^{(n)}}{\sum_k^C \mathcal{G}\left(x_i,\mu_k^{(n)},\sigma_k^{(n)}\right)\pi_k^{(n)}} \qquad \pi_j^{(n+1)} = \frac{1}{n}\sum_{i=1}^{n} h_{ij}$$

$$\mu_j^{(n+1)} = \frac{\sum_i^n h_{ij}x_i}{\sum_i^n h_{ij}} \qquad \sigma_j^{2(n+1)} = \frac{\sum_i^n h_{ij}(x_i-\mu_j)^2}{\sum_i^n h_{ij}}$$

MAP-EM

1. **E-step**: compute

$$E_{Z|X,\Psi}[\log P_{\Psi|X,Z}(\Psi|\mathcal{D},z)\,|\,\mathcal{D},\Psi^{(n)}] \Rightarrow Q(\Psi|\Psi^{(n)}) + \log P_\Psi(\Psi) \quad \text{(only need to compute } Q\text{)}$$

2. **M-step**: compute

$$\Psi^{(n+1)} = \arg\max_{\Psi}\{Q(\Psi|\Psi^{(n)}) + \log P_\Psi(\Psi)\}.$$