Chapter 2: Optimal Classification

- **Error of classifier.**: $\epsilon[\psi(x)] = P(\psi(X) \neq Y) = \underbrace{p(\psi(X)=1|Y=0)}_{\epsilon^0=\int_{\{x|\psi(x)=1\}} p(x|Y=0)dx} P(Y=0) + \underbrace{p(\psi(X)=0|Y=1)}_{\epsilon^1=\int_{\{x|\psi(x)=0\}} p(x|Y=1)dx} P(Y=1)$
- **Cond. error**: $\epsilon[\psi|X] = P(\psi(X) \neq Y|X=x) = P(\psi(X)=0, Y=1|X=x) + P(\psi(X)=1, Y=0|X=x) = I_{\{\psi(x)=0\}}\eta(x) + I_{\{\psi(x)=1\}}(1-\eta(x))$
- **Post.prob.func.**: $\eta(x) = E[Y|X=x] = P(Y=1|X=x)$
- **Sensitivity**: $1-\epsilon^1[\psi]$; **Specificity**: $1-\epsilon^0[\psi]$
- **Thm.** *Bayes classifier*:

$$\psi^*(x) = \arg\max_i P(Y=i|X=x) = \begin{cases}1, & \eta(x) > \frac{1}{2}\\ 0, & \text{otherwise}\end{cases} \quad (1)$$

- **Thm. Bayes Error**: $\epsilon^* = P(Y=0)\epsilon^0[\psi^*] + P(Y=1)\epsilon^1[\psi^*] = E[\min\{\eta(X), 1-\eta(x)\}] = \frac{1}{2} - \frac{1}{2}E[|2\eta(X)-1|]$
- **Bayes class.**: $\psi^*(x) = \begin{cases}1 & \overbrace{D^*(x)}^{\text{opt. discriminant}} > \overbrace{k^*}^{\text{opt. threshold}} \\ & = P(Y=1)p(x|Y=1) > P(Y=0)p(x|Y=0)\\ 0, & \text{otherwise}\end{cases}$
- $D^*(x) = \ln\frac{p(x|Y=1)}{p(x|Y=0)}$; $k^* = \ln\frac{P(Y=0)}{P(Y=1)}$

Gaussian Prob.: $p(x|Y=i) = \frac{1}{\sqrt{(2\pi)^d\det(\Sigma_i)}}\exp[\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)]$

- $D^*(x) = \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \frac{1}{2}\ln\frac{\det(\Sigma_0)}{\det(\Sigma_1)}$

Homo. Case: Let $\|x_0 - x_1\|_\Sigma = \sqrt{(x_0-x_1)^T\Sigma^{-1}(x_0-x_1)}$

$$\psi_L^*(x) = \begin{cases}1, & \|x-\mu_1\|^2_\Sigma < \|x-\mu_0\|^2_\Sigma + 2\ln\frac{P(Y=1)}{P(Y=0)}\\ & = a^Tx + b > 0\\ 0, & \text{otherwise}\end{cases}$$

- $a = \Sigma^{-1}(\mu_1 - \mu_0)$
- $b = (\mu_0-\mu_1)^T\Sigma^{-1}(\frac{\mu_0+\mu_1}{2}) + \ln\frac{P(Y=1)}{P(Y=0)}$
- $\epsilon_L^* = c\Phi(\frac{k^*-\frac{1}{2}\delta^2}{\delta}) + (1-c)\Phi(\frac{-k^*-\frac{1}{2}\delta^2}{\delta}), \delta = \sqrt{(\mu_1-\mu_0)^T\Sigma^{-1}(\mu_1-\mu_0)}$

Heter. Case: $\psi_Q^*(x) = \begin{cases}1, & x^TAx + b^Tx + c > 0,\\ 0, & \text{otherwise}\end{cases}$

- $A = \frac{1}{2}(\Sigma_0^{-1} - \Sigma_1^{-1})$
- $b = \Sigma_1^{-1}\mu_1 - \Sigma_0^{-1}\mu_0$
- $c = \frac{1}{2}(\mu_0^T\Sigma_0^{-1}\mu_0 - \mu_1^T\Sigma_1^{-1}\mu_1) + \frac{1}{2}\ln\frac{\det\Sigma_0}{\det\Sigma_1} + \ln\frac{P(Y=1)}{P(Y=0)}$

Chapter 3: Sample-Based Classification

- **No-Free-Lunch**: One can never know if their finite-sample performance will be satisfactory, no matter how large n is. $E[\epsilon_n] \geq \frac{1}{2} - \tau$

Chapter 4: Parametric Classification

LDA — Homo. Gaussian Case

- **Linear Discriminant Analysis (LDA)**: $\hat{\Sigma}_0^{ML} = \frac{1}{N_0-1}\sum_{i=1}^n(X_i-\hat{\mu}_0)(X_i-\hat{\mu}_0)^TI_{Y_i=0}$, $\hat{\Sigma} = \frac{\hat{\Sigma}_0+\hat{\Sigma}_1}{2}$
  - Boundary: $a_n^Tx + b_n = k_n$.
    - $a_n = \hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_0) = \begin{bmatrix}a_1\\a_2\end{bmatrix}$
    - $b_n = (\hat{\mu}_0-\hat{\mu}_1)^T\hat{\Sigma}^{-1}(\frac{\hat{\mu}_0+\hat{\mu}_1}{2}) = number$
- **Diagnoal LDA**: Make $\hat{\Sigma} \to \hat{\Sigma}_D = \begin{bmatrix}\Sigma_{1,1} & 0\\ 0 & \Sigma_{2,2}\end{bmatrix}$
- **Nearest-Mean Class.(NMC)**: $\hat{\Sigma}_M = \begin{bmatrix}\hat{\sigma}^2_{ij} & 0\\ 0 & \hat{\sigma}^2_{ij}\end{bmatrix}$. $\hat{\sigma}^2 = \sum_{k=1}^d(\hat{\Sigma})_{kk}$. Given $k_n = 0$, $a = \hat{\mu}_1 - \hat{\mu}_0$ $b = (\hat{\mu}_0 - \hat{\mu}_1)^T\left(\frac{\hat{\mu}_0+\hat{\mu}_1}{2}\right)$. Boundary is $\perp$ means
- 2D: $a_1x_1 + a_2x_2 + b_n = 0$
- **Logistic Class.**: linear classification
  - $logit(\eta(x|a,b)) = \ln(\frac{\eta(x|a,b)}{1-\eta(x|a,b)}) = a^Tx + b$
  - $L(a,b|S_n) = \ln\left(\prod_{i=1}^n P(Y=Y_i|X=X_i)\right) = \sum_{i=1}^n \ln(\eta(X_i|a,b)^{Y_i}(1-\eta(X_i|a,b))^{1-Y_i})$
- LDA Classifier: $\psi_n(x)\begin{cases}1, & a_n^T + b_n > 0\\ 0, & \text{otherwise}\end{cases}$
- $\epsilon_n = (1-c)\Phi\left(\frac{a_n^T\mu_0+b_n}{\sqrt{a_n^T\Sigma_0a_n}}\right) + c\Phi\left(-\frac{a_n^T\mu_1+b_n}{\sqrt{a_n^T\Sigma_1a_n}}\right)$

QDA — Heter. Gaussian Case

- **Boundry**: $x^TA_nx + b_n^Tx + c + n = k_n$
  - $A_n = -\frac{1}{2}(\hat{\Sigma}_1^{-1} - \hat{\Sigma}_0^{-1}) = \begin{bmatrix}a_{11} & a_{12}\\ a_{12} & a_{22}\end{bmatrix}$
  - $b_n = \hat{\Sigma}_1^{-1}\hat{\mu}_1 - \hat{\Sigma}_0^{-1}\hat{\mu}_0 = \begin{bmatrix}b_1\\b_2\end{bmatrix}$
  - $c_n = -\frac{1}{2}(\hat{\mu}_1^T\hat{\Sigma}_1^{-1}\hat{\mu}_1 - \hat{\mu}_0^T\hat{\Sigma}_0^{-1}\hat{\mu}_0) - (\frac{1}{2}\ln\frac{|\hat{\Sigma}_1|}{|\hat{\Sigma}_0|}) = number$
- 2D: $a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 + b_1x_1 + b_2x_2 + c = 0$

Chapter 5:

- $\eta_{n,h}(x) = \sum_{i=1}^n W_{n,h}(x,X_i)I_{Y_i=1}$.
- **Weights**: $W_{n,h}(x,X_i) \geq 0$; $\sum_{i=1}^n W_{n,h}(x,X_i) = 1$
- **Plug-in classifier**: $\psi_n(x) = \begin{cases}1, & \sum_{i=1}^n W_{n,h}(x,X_i)I_{Y_i=1} > \sum_{i=1}^h W_{n,h}(x,X_i)I_{Y_i=0}\\ 0, & \text{otherwise}\end{cases}$
- **Histogram Class.**: $W_{n,h}(x,X_i) = \begin{cases}\frac{1}{N_h(x)}, & X_i \in A_h(x)\\ 0, & \text{otherwise}\end{cases}$
- **Kernel Class.**: $W_{n,h}(x,X_j) = \frac{k(\frac{x-X_j}{h})}{\sum_{i=d1}^n k(\frac{x-X_i}{h})}$. h is the kernel bandwidth (smoothing parameter). Small $h \to$ overfitting
- **Thm. Cover-Hart**: $\epsilon_{NN} = E[2\eta(X)(1-\eta(x))]$
- $\epsilon_{kNN} = E[\alpha_k(\eta(X))]$.
- $\alpha_k(p) = \sum_{i=1}^{(k-1)/2}\binom{k}{i}p^{i+1}(1-p)^{k-1} + \sum_{i=(k+1)/2}^k\binom{k}{i}p^i(1-p)^{k+1-i}$
- Find p_0 s.t. $a_k = \alpha_k'(p_0) = \frac{\alpha_k(p_0)}{p_0}$. $a_k > 1$, $p \in [0, \frac{1}{2}]$
- **Thm. Asymptotic class. error of NN**: $\epsilon_{NN} = \begin{cases}2\epsilon^*(1-\epsilon^*) \text{ iff } \eta(X) \in \{\epsilon^*, 1-\epsilon^* *\}\\ \epsilon^* \text{ iff } \eta(X) \in \{0, \frac{1}{2}, 1\}\end{cases}$

- **Stone's Thm**: The class. rule is universally consistent, if
  1. $\sum_{i=1}^n W_{n,i}(X)I_{\|X_i-X\|>\delta} \to^P 0$, as $n \to \infty$, for all $\delta > 0$
  2. $\max_{i=1,\dots,n} W_{n,i}(X) \to^p 0$, as $n \to \infty$
  3. There ia a constant $c \geq 1$ such that , for every nonnegative $f: R^d \to R$, and all $n \geq 1$, $E[\sum_{i=1}^n W_{n,i}(X)f(X_i)] \leq cf(X)$
- **Uni. Consist. of Histrogram Class.**:
  - $diam[A_n(X)] = \sup_{x,y\in A_n(X)}\|x-y\| \to 0$ in probability.
  - $N_n(X) \to \infty$
- **Uni. Consist. of Cubic Histogram**: Let $V_n = h_n^d$. If $h_n \to 0$,but $nV_n \to \infty$ as $n \to \infty$. Then $E[\epsilon_n] \to \epsilon^*$
- **Uni. Consist. of kNN**: If $K \to \infty$ while $\frac{K}{n} \to 0$ as $n \to \infty$. Then $E[\epsilon_n] \to \epsilon^*$.
- **Uni. Consist. of Kernel**: $h_n \to 0$ with $nh_n^d \to \infty$ as $n \to \infty$. (kernel k is nonnegative, cont. integrable)

Key points & Definitions

- The posterior probability function is needed to define the Bayes classifier.; Bayes error is optimal error; LDA is parameteric.
- *minimum and the maximal of the Bayes error of binary classification*: $\epsilon^* = E[\min\{\eta(X), 1-\eta(X)\}]$.
- *expected classification error $\mu = E[error_n]$ not a function of the training data?*: μ_n is data-independent, it is a function only of the classification rule.
- *meaning of an error estimator is optimistically biased?*: Be significantly smaller on average than the true error, due to overfitting. When the bias < 0, and left shifted.
- *Is a consistent classification rule always better than a non-consistent one and why?*: No. non-consist. is better when n is small because consist. class. tned to be complex.
- *If a classifier is overfitted, will its apparent error?*: Apparent error is smaller due to small sample size.
- *The penalty term in an SVM?*: Small C includes outlier (soft margin and less overfitting)
- *Cover-Hart Thm.*: The expected error of the NN classification rule satisfies $\epsilon_{NN} = \lim_{n\to\infty}E[\epsilon_n] = E[2\eta(X)(1-\eta(X))]$. $\epsilon_{NN} \leq 2\epsilon^*(1-\epsilon^*) \leq 2\epsilon^*$. "The error of the nearest-neighbot classifier with a large sample size cannot be worse than two times the Bayes error." $\epsilon_{NN} \geq \epsilon_{3NN} \geq \epsilon^*$
- **Ch. 1**: *Curse of dimen. (peaking phen.)*: With fixed sample size, class. error improve with more features, then decreases.
- *Scissors Effect*: Simpler classification rules can perform better under small sample sizes. On the contray in big data.
- **Ch. 3**: *Classification rule vs. classifier*: output classifiers ; class lables. *Consistency*: As $n \to \infty$, $\epsilon_n \to \epsilon^*$
- **Ch. 5**: *Nonparametric class.* has no assumption about the shape of the distributions. use smoothing. Selecting right amount of smoothing given n and complexity of dist. *Weights*: adding the influences of each data point (X_i, Y_i). *3/5NN* rule are better than 1NN under small sample size

Math facts

- Bayes: $P(Y=0|X=x) = \frac{P(Y=0)P(x|Y=0)}{P(x)}$
- $\det\begin{bmatrix}a & b\\ c & d\end{bmatrix} = ad - bc$; $\begin{bmatrix}a & b\\ c & d\end{bmatrix}^{-1} = \frac{1}{ad-bc}\begin{bmatrix}d & -b\\ -c & a\end{bmatrix}$
- Affine trans. $f(x) = AX + B$. If $X \sim N(\mu, \Sigma)$, $a^TX + b \sim N(A^T\mu + b, A^T\Sigma A)$.
- *Convergence in prob.*: $X_n \to^P X$. $\lim_{n\to\infty}P(|X_n - X| > \tau) = 0$, for all $\tau > 0$. Implies that $f(X_n) \to^P f(X)$
- *Gauss. CDF*: $1 - \Phi(-a) = \Phi(a)$