Introduction to Machine Learning and Pattern Recognition

David Brady[1]

[1]ECE Department
Northeastern University

Fall 2018

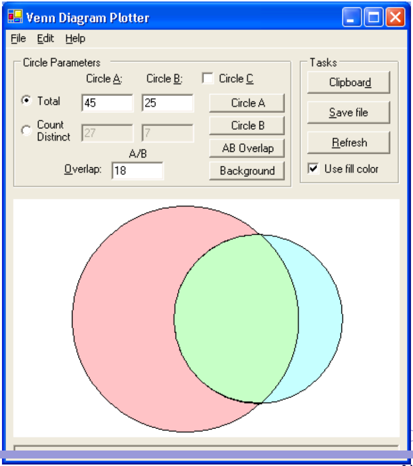
Next:

a Priori Probability

How do we use *a priori* information in classifying sea bass or salmon?
What if we know that 90% of fish are bass and 10% are salmon?
How does the best classifier change if 50% are bass and 50% are salmon?

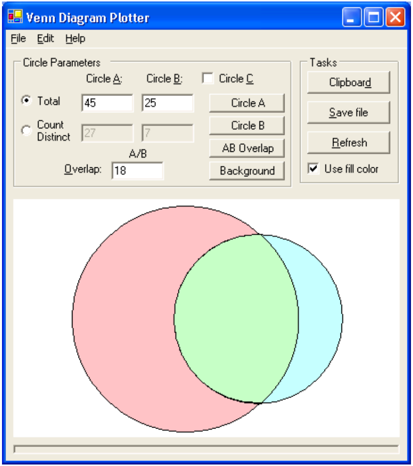
Some notation:

- x is a column vector of new features
- x is a point in plot for each "experiment"
- *a priori* = "before the observation"
- ω is an experimental outcome (new feature)
  - $\omega = \omega_1 \mapsto$ x in right circle
  - $\omega = \omega_2 \mapsto$ x in left circle
- *a prior* probabilities
  - $P[\omega = \omega_1]$ = Probability x falls in right circle
  - $P[\omega = \omega_2]$ = Probability x falls in left circle
- $P[\omega = \omega_1] < P[\omega = \omega_2]$
- $P[\omega_1]$ is not related to area of right circle



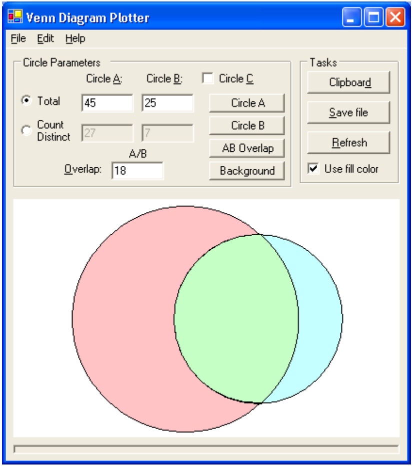
A Posteriori Probability

- *a posteriori* = "after the observation"
- *a posteriori* probabilities
  - $P[\omega_1|\mathbf{x}]$ = chance $\mathbf{x}$ in right circle after observing $\mathbf{x}$
  - $P[\omega_2|\mathbf{x}]$ = chance $\mathbf{x}$ in left circle after observing $\mathbf{x}$
- $P[\omega_2] > P[\omega_1]$ *(a priori)*
- $P[\omega_2|\mathbf{x} \textit{ in green}]$? $P[\omega_1|\mathbf{x} \textit{ in green}]$ *(a posteriori)*
- $P[\omega_2|\mathbf{x} \textit{ in red}]$? $P[\omega_1|\mathbf{x} \textit{ in red}]$ *(a posteriori)*
- $P[\omega_2|\mathbf{x} \textit{ in blue}]$? $P[\omega_1|\mathbf{x} \textit{ in blue}]$ *(a posteriori)*



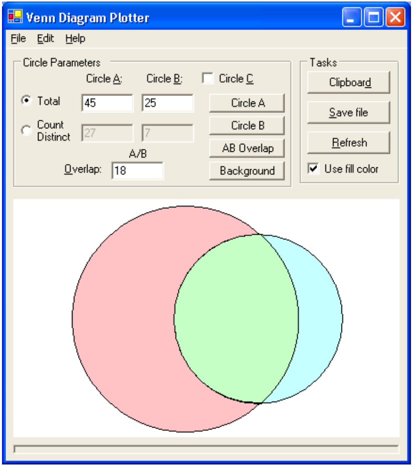
Likelihood

- $p[\mathbf{x}|\omega_1]$ is called the *likelihood of ω_1 with respect to* $\mathbf{x}$
- for a continuous random vector $\mathbf{x}$, $p[\mathbf{x}|\omega_1]$ is the feature vector probability density function if $\omega = \omega_1$
  - $\int_{\mathbf{x}\in\mathscr{X}} p\left[\mathbf{x}|\omega_j\right] d\mathbf{x} = 1$
- for a discrete random vector $\mathbf{x}$, $p[\mathbf{x}|\omega_j]$ is the feature vector probability mass function if $\omega = \omega_j$
  - $\sum_{\mathbf{x}\in\mathscr{X}} p\left[\mathbf{x}|\omega_j\right] = 1$



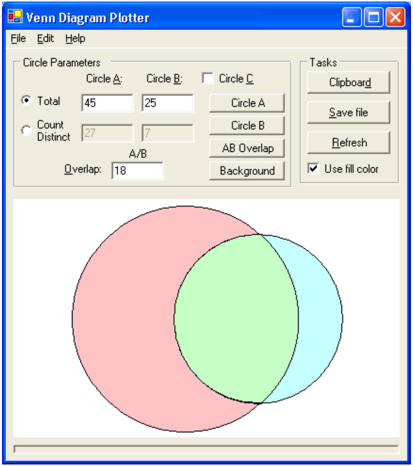
Likelihood (2)

- $p[\mathbf{x}|\omega_1] = 0$ if $\mathbf{x}$ outside of the right circle
- special case: uniform distributions
  - $p[\mathbf{x}|\omega_1] = 1/$ area of right circle
  - $p[\mathbf{x}|\omega_2] = 1/$ area of left circle
- for $\mathbf{x}$ in the green area, $p[\mathbf{x}|\omega_1] > p[\mathbf{x}|\omega_2]$ (we say that ω_1 is more likely there)
- so, if $\mathbf{x}$ falls into the green area, which ω_j would you choose?



Bayes Formula

- $p[\mathsf{x}]$ is the *evidence* of x
- $$p[\mathsf{x}] = \sum_j P[\omega_j] p[\mathsf{x}|\omega_j]$$
- $p[\mathsf{x}]$ is the pdf (or pmf) of x
- Bayes Formula: *posterior* = *likelihood* · *prior* / *evidence*
- $$P[\omega_j|\mathsf{x}] = p[\mathsf{x}|\omega_j] \cdot P[\omega_j] / p[\mathsf{x}]$$



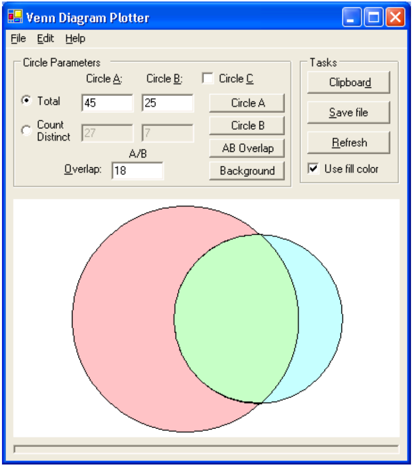
Next:

Bayes Decision Rule

- so, if **x** falls into the green area, which ω_j would you choose?
- Bayes' Decision Rule for Uniform Costs: choose j to maximize $P[\omega_j|\mathbf{x}]$

special case: uniform distribution, uniform costs

- choose ω_1 if $P[\omega_1|\mathbf{x}] = p[\mathbf{x}|\omega_1] \cdot P[\omega_1]/p[\mathbf{x}] > p[\mathbf{x}|\omega_2] \cdot P[\omega_1]/p[\mathbf{x}] = P[\omega_2|\mathbf{x}]$
- $p[\mathbf{x}|\omega_1] \cdot P[\omega_1] > p[\mathbf{x}|\omega_2] \cdot P[\omega_1]$
- three cases:
- area of left circle / area of right circle $> P[\omega_1]/P[\omega_2]$

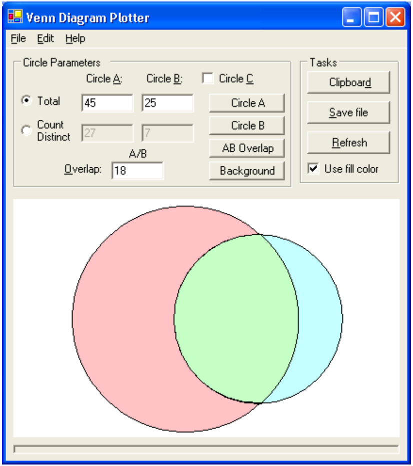


Bayes Decision Rule (2)

- Bayes' Decision Rule for Uniform Costs: choose j to maximize $P[\omega_j|\mathbf{x}]$

special case: uniform distribution, uniform costs

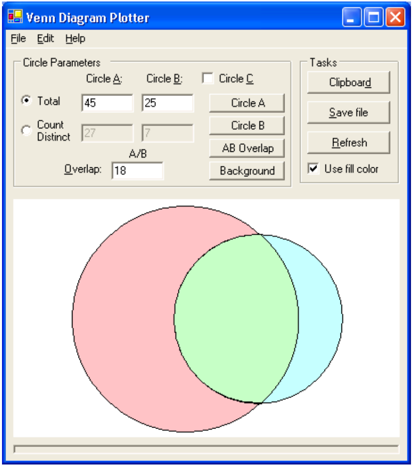
- choose ω_1 if $P[\omega_1|\mathbf{x}] = p[\mathbf{x}|\omega_1] \cdot P[\omega_1]/p[\mathbf{x}] > p[\mathbf{x}|\omega_2] \cdot P[\omega_2]/p[\mathbf{x}] = P[\omega_2|\mathbf{x}]$
- if $p[\mathbf{x}|\omega_1] \cdot P[\omega_1] > p[\mathbf{x}|\omega_2] \cdot P[\omega_2]$, decide $\omega = \omega_1$
- what to do with "=" ? (later)
  - (think of each side as functions of $\mathbf{x}$, which is observed)



Bayes Decision Rule (3)

- Bayes' Decision Rule says choose j to maximize $P[\omega_j|\mathbf{x}]$
- 4 cases:

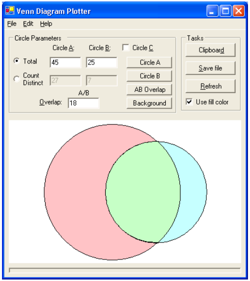
1. $\mathbf{x}$ falls in red area
   1.1 $0 \cdot P[\omega_1] >$? $P[\omega_2]$/*area of left circle* $\longleftarrow$ *never*!
   1.2 decide $\omega = \omega_2$
2. $\mathbf{x}$ falls in blue area
   2.1 $P[\omega_1]$/*area of right circle* $>$?$0 \cdot P[\omega_2]$ $\longleftarrow$always!
   2.2 decide $\omega = \omega_1$



Bayes Decision Rule (4)

- Bayes' Decision Rule : choose j to maximize $P[\omega_j|\mathbf{x}]$
- 2 more cases:

  3. $\mathbf{x}$ falls in white area
     1. $0 \cdot P[\omega_1] = \ P[\omega_1] \cdot 0$ *tie*!
     2. decide either $\omega = \omega_1$ *or* ω_2
  4. $\mathbf{x}$ falls in green area
     1. $P[\omega_1]/$*area of* right! *left* *circle* $>$? $P[\omega_2]/$*area of* *right* left! $\leftarrow$depends on priors!
     2. decide $\omega = \omega_1$ if above inequality is satisfied



Next:

Action Function, Conditional Loss, Conditional Risk, Average Risk

- we design *action function*
- say $\omega_1 \Longleftrightarrow$ action function $\alpha(\mathbf{x}) = \alpha_1$
- $\alpha(\mathbf{x})$ depends only on $\mathbf{x}$

$\lambda(\alpha_i, |\omega_j)$ is the *conditional loss* of action $\alpha(\mathbf{x}) = \alpha_i$ when ω_j occurs

$\lambda(\alpha_i\|\omega_j)$	ω_1	ω_2
α_1	0USD	43USD
α_2	9700USD	0USD

- some mistakes are more expensive

- Conditional Risk is $R(\alpha(\mathbf{x})|\mathbf{x})$
- depends on observation and action rule
- $R(\alpha(\mathbf{x}) = \alpha_i|\mathbf{x}) = \sum_j \lambda(\alpha_i|\omega_j)P[\omega_j|\mathbf{x}]$
- average risk $R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x} = E\left[R(\alpha(\mathbf{x})|\mathbf{x})\right]$
- Bayes Decision Rule $\alpha_{Bayes}(\mathbf{x})$ minimizes *average risk*

Minimum Risk Solutions

- special case: two classes ω_1 and ω_2
- two actions: α_1 and α_2
- Bayes classifiers minimize conditional risk
- $R(\alpha_i|\mathbf{x}) = \lambda(\alpha_i|\omega_1)P[\omega_1|\mathbf{x}] + \lambda(\alpha_i|\omega_2)P[\omega_2|\mathbf{x}]$
- $for\ each\ \mathbf{x},\ \alpha_{Bayes}(\mathbf{x}) = argminR(\alpha_i|\mathbf{x})$
- say ω_1 if: $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$, or
- $\lambda(\alpha_1|\omega_1)P[\omega_1|\mathbf{x}] + \lambda(\alpha_1|\omega_2)P[\omega_2|\mathbf{x}] <$ $\lambda(\alpha_2|\omega_1)P[\omega_1|\mathbf{x}] + \lambda(\alpha_2|\omega_2)P[\omega_2|\mathbf{x}]$
- $P[\omega_1|\mathbf{x}](\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1)) > P[\omega_2|\mathbf{x}](\lambda(\alpha_1|\omega_2) - \lambda(\alpha_2|\omega_2))$
- $\frac{P[\omega_1|\mathbf{x}]}{P[\omega_2|\mathbf{x}]} > \frac{\lambda(\alpha_1|\omega_2)-\lambda(\alpha_2|\omega_2)}{\lambda(\alpha_2|\omega_1)-\lambda(\alpha_1|\omega_1)}$
- $\frac{P[\mathbf{x}|\omega_1]}{P[\mathbf{x}|\omega_2]} > \frac{P[\omega_2]}{P[\omega_1]}\frac{\lambda(\alpha_1|\omega_2)-\lambda(\alpha_2|\omega_2)}{\lambda(\alpha_2|\omega_1)-\lambda(\alpha_1|\omega_1)} \Longrightarrow$ say ω_1
- $\frac{P[\mathbf{x}|\omega_1]}{P[\mathbf{x}|\omega_2]} = \frac{P[\omega_2]}{P[\omega_1]}\frac{\lambda(\alpha_1|\omega_2)-\lambda(\alpha_2|\omega_2)}{\lambda(\alpha_2|\omega_1)-\lambda(\alpha_1|\omega_1)} \Longrightarrow$ say ω_1 with prob. p

Randomization

- randomization:
- $\frac{P[\mathbf{x}|\omega_1]}{P[\mathbf{x}|\omega_2]} = \frac{P[\omega_2]}{P[\omega_1]} \frac{\lambda(\alpha_1|\omega_2) - \lambda(\alpha_2|\omega_2)}{\lambda(\alpha_2|\omega_1) - \lambda(\alpha_1|\omega_1)} \Longrightarrow$ say ω_1 with prob. p
- all values of p produce the same average risk, so $p = 0, 1$ is fine (no randomization)
- p is sometime used to ease calculations (later)

Interpretation of Bayes Decision Rule

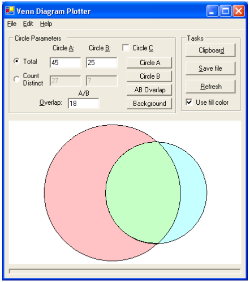
- $\frac{P[\mathbf{x}|\omega_1]}{P[\mathbf{x}|\omega_2]} > \frac{P[\omega_2]}{P[\omega_1]} \frac{\lambda(\alpha_1|\omega_2)-\lambda(\alpha_2|\omega_2)}{\lambda(\alpha_2|\omega_1)-\lambda(\alpha_1|\omega_1)}$
- $\frac{P[\mathbf{x}|\omega_1]}{P[\mathbf{x}|\omega_2]}$ is the likelihood ratio
- $\frac{P[\omega_2]}{P[\omega_1]}$ is the ratio of a priori probabilities
- $\frac{\lambda(\alpha_1|\omega_2)-\lambda(\alpha_2|\omega_2)}{\lambda(\alpha_2|\omega_1)-\lambda(\alpha_1|\omega_1)}$ is the ratio of relative costs
- if the likelihood ratio is sufficiently large, accept ω_1
- if the ratio of prior probabilities is sufficiently large, reject ω_1
- if the relative cost of accepting ω_1 is sufficiently large, reject ω_1

Example of Bayes Decision Rule

- right circle (ω_1) has area = 25
- left circle (ω_2) has area = 45
- overlap has area=18
- uniform distribution on circle for ω_j

$\lambda(\alpha_i\|\omega_j)$	ω_1	ω_2
α_1	0	1
α_2	1	0

uniform costs



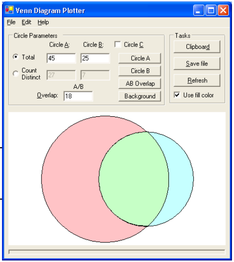
- ratio of relative costs = 1
- likelihood ratio($\mathbf{x} \in$ red) = 0
- likelihood ratio($\mathbf{x} \in$ blue) = ∞
- likelihood ratio($\mathbf{x} \in$ white) undefined (why?)

Example of Bayes Decision Rule (cont'd)

- $\frac{P[\omega_2]}{P[\omega_1]}$ is variable
- Bayes Decision Rule:
- $\frac{P[\mathbf{x}|\omega_1]}{P[\mathbf{x}|\omega_2]} > \frac{P[\omega_2]}{P[\omega_1]} \frac{\lambda(\alpha_1|\omega_2)-\lambda(\alpha_2|\omega_2)}{\lambda(\alpha_2|\omega_1)-\lambda(\alpha_1|\omega_1)} \Longrightarrow$ say ω_1
-

$$\boxed{\frac{P[\mathbf{x}|\omega_1]}{P[\mathbf{x}|\omega_2]} > \frac{P[\omega_2]}{P[\omega_1]} \cdot 1 \Longrightarrow \text{say } \omega_1}$$

- $\mathbf{x} \in$ red, accept ω_2
- $\mathbf{x} \in$ blue, accept ω_1
- $\mathbf{x} \in$ white, undefined
- $\mathbf{x} \in$ green, $45/25 \overset{?}{>} \frac{P[\omega_2]}{P[\omega_1]}$. If so, accept ω_1.
- $\mathbf{x} \in$ green, $45/25 \overset{?}{=} \frac{P[\omega_2]}{P[\omega_1]}$. If so, accept ω_1 with probability p



- let $P[\omega_1] = P[\omega_2] = 0.5$, so $\frac{P[\omega_2]}{P[\omega_1]} = 1$.
- randomization not needed
- if ω_2 is sufficiently rare, accept ω_1 here

Conditional Risk For Uniform Costs

conditional risk: confusing

$$r(\alpha_i|\underline{x})\ \cancel{R_i} = R(\alpha(\mathbf{x}) = \alpha_i|\mathbf{x}) = \sum_j \lambda(\alpha_i|\omega_j)P[\omega_j|\mathbf{x}]$$

better

uniform costs: $\lambda(\alpha_i|\omega_j) = 1 - \delta_{ij} = \begin{cases} 1, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases}$

conditional risk for uniform costs:

$$r(\alpha_i|\underline{x})\ \cancel{R_i} = \sum_{j \neq i} P[\omega_j|\mathbf{x}] = P[error|\mathbf{x}]$$

correct

$$= 1 - P[\omega_i|\underline{x}]$$

also good

For uniform costs, Bayes Decision Rule minimizes the error probability for each $\mathbf{x}$, $P[error|\mathbf{x}]$, and the average error probability $R = E[P[error|\mathbf{x}]] = P[error]$.

Computing Average Risk

average risk:

$\alpha(\underline{x})$ = our decision at $\underline{x}$
$R(\alpha(\underline{x})|\underline{x})$ is the risk of that decision

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x} = E\left[R(\alpha(\mathbf{x})|\mathbf{x})\right]$$

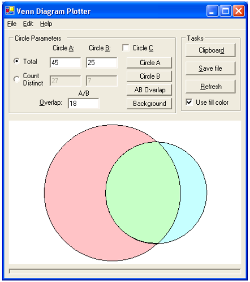
$p(\mathbf{x}) = \sum_j P(\omega_j)p(\mathbf{x}|\omega_j)$, so one way to compute average risk is by conditioning on ω_i first, then average over a priori probabilities: First find

$$R_j = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x}|\omega_j)d\mathbf{x}$$

$$\textit{then } R = \sum_j P(\omega_j)R_j$$

where R_j is the conditional risk ~~if~~ given ω_j occurs
There are other ways! (homework)

Example: Computing Average Risk



- in summary, our Bayes Decision Rule is:
- α_1 if **x** in blue or green
- α_2 if **x** in red
- $R = P\,[error]$ for uniform costs
- First find risk under under each class, R_i
- $R_1 = P\,[error|\omega_1] = 0$
- $R_2 = P\,[error|\omega_2] = \boxed{P[\omega_2]}P[\mathbf{x} \in green|\omega_2]$
- average over a priori probabilities
- $P\,[error] = 0.5 \cdot 0 + 0.5\frac{18}{45} = 0.20$

Next:

If the *a priori* probabilities are not known?

- Consider two-class case, and suppose $P[\omega_1]$ is unknown
- Let $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$
- Let $\mathscr{R}_i$ be the decision region for class ω_i
- Rewrite R_i

$$R_2 = \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathscr{R}_1} p(\mathbf{x}|\omega_2) d\mathbf{x}$$

$$R_1 = \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{\mathscr{R}_2} p(\mathbf{x}|\omega_1) d\mathbf{x}$$

$$R = P[\omega_1]R_1 + P[\omega_2]R_2 = R_2 + P[\omega_1](R_1 - R_2)$$

- want risk to be independent of unknown $P[\omega_1] \Longrightarrow$ find $\mathscr{R}_i$ so that $R_1 = R_2$
- then $R_{minmax} = R_1 = R_2$

Finding the Minimax Bayes Decision Rule

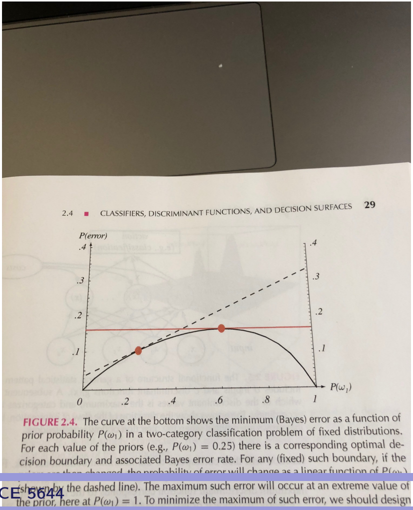
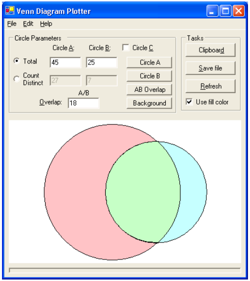

2.4 ■ CLASSIFIERS, DISCRIMINANT FUNCTIONS, AND DECISION SURFACES 29

FIGURE 2.4. The curve at the bottom shows the minimum (Bayes) error as a function of prior probability $P(\omega_1)$ in a two-category classification problem of fixed distributions. For each value of the priors (e.g., $P(\omega_1) = 0.25$) there is a corresponding optimal decision boundary and associated Bayes error rate. For any (fixed) such boundary, if the priors are then changed, the probability of error will change as a linear function of $P(\omega_1)$ (shown by the dashed line). The maximum such error will occur at an extreme value of the prior, here at $P(\omega_1) = 1$. To minimize the maximum of such error, we should design

- Note that when $R_1 = R_2$, $\frac{\delta R}{\delta P[\omega_1]} = 0$
- R_{Bayes} is maximized at the minmax Bayes Decision Rule
- 2 ways to find the minimax decision rule:
  - fix $P[\omega_1]$, and find R_i for the Bayes Decision Rule ($\mathscr{R}_i$).Then either:

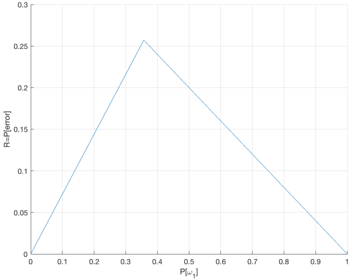
    1. choose the decision rule $(P[\omega_1], \mathscr{R}_1)$ so that $R_1 = R_2$, or
    2. choose the decision rule so that $R = P[\omega_1](R_1 - R_2) + R_2$ is maximum

Example: Minimax Bayes Decision Rule



- $R = P[\omega_1]R_1 + P[\omega_2]R_2 = P[error]$
- $\mathbf{x} \in$ green, accept ω_1 if $\frac{45}{25} > \frac{P[\omega_2]}{P[\omega_1]}$. (if equality, flip biased coin to accept!)
- $R_1 = \begin{cases} \frac{18}{25}, & P[\omega_1] < \frac{5}{14} \\ 0, & P[\omega_1] > \frac{5}{14}. \end{cases}$
- $R_2 = \begin{cases} 0 & P[\omega_1] < \frac{5}{14} \\ \frac{18}{45}, & P[\omega_1] > \frac{5}{14}. \end{cases}$
- $R = \begin{cases} P[\omega_1]\frac{18}{25} + 0, & P[\omega_1] < \frac{5}{14} \\ 0 + \frac{18}{45}(1 - P[\omega_1]), & P[\omega_1] > \frac{5}{14} \end{cases}$

Example: Minmax Bayes Decision Rule (2)



- worst-case risk R= 9/35 at $P[\omega_1] = 5/14$
- minmax Bayes detector is Bayes Decision Rule at $P[\omega_1] = 5/14$.
- $\mathbf{x} \in$ green, accept ω_1 with probability p (flip coin)
- find p: $R_1 = (1-p)\,\frac{18}{25}$; $R_2 = p\frac{18}{45}$
- $R_1 = R_2$ or $R = R_2$ implies $p = \frac{9}{14}$

Next:

What to do if costs and priors are unknown?

- consider the j-class case
- let ω_1 correspond to the *null* class
- consider all decision rules satisfying a (false-alarm, level) constraint: $R_1 \leq constant$
- create a detector which minimizes the residual risk: $min \ \sum_{j \neq 1} R_j$
- for j=2, the NP-optimal detector is a likelihood-ratio test

$$\frac{p(\mathrm{x}|\omega_2)}{p(\mathrm{x}|\omega_1)} > t, \ accept \ \omega_2$$

$$\frac{p(\mathrm{x}|\omega_2)}{p(\mathrm{x}|\omega_1)} = t, \ accept \ \omega_2 \ with \ prob \ p$$

- threshold t and probability p are set by the above (false-alarm, level) constraint
- same structure as Bayes Decision Rule, different threshold

Next:

Discriminant Functions and Classifiers

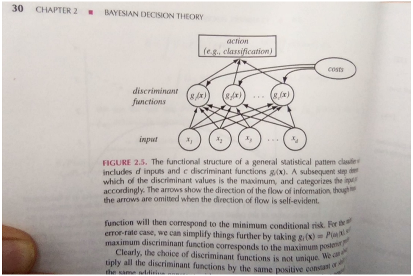

30 CHAPTER 2 ■ BAYESIAN DECISION THEORY

FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident.

function will then correspond to the minimum conditional risk. For the minimum-error-rate case, we can simplify things further by taking $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$, so that the maximum discriminant function corresponds to the maximum posterior probability.

Clearly, the choice of discriminant functions is not unique. We can always multiply all the discriminant functions by the same positive constant or shift them by the same additive constant

- a *discriminant function* $g_i(\cdot)$ is a functional mapping a feature $\mathbf{x}$ to a measure of fit for class ω_i
- a *classifier* assigns each feature $\mathbf{x}$ to one of c classes using c discriminator function comparisons
- $\mathbf{x}$ is assigned to class i if $i = argmax_j g_j(\mathbf{x})$
- Bayes classifier uses $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$

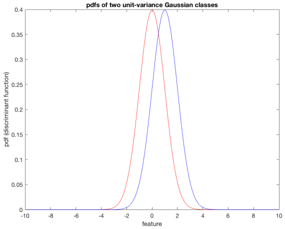
Faster Code ⇔ Simplify Discriminant Functions!

$$\begin{aligned} argmax_j g_j(\mathbf{x}) &= argmax_j \; log\,(g_j(\mathbf{x})) \\ &= argmax_j \; exp\,(g_j(\mathbf{x})) \\ &= argmax_j \; 43 \cdot (g_j(\mathbf{x})) \end{aligned}$$

- coding and analysis are eased by simplifying comparisons
- the same Bayes classifier can use any of the following sets:

$$\begin{aligned} g_i(\mathbf{x}) &= \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)} \\ &= p(\mathbf{x}|\omega_i)P(\omega_i) \\ &= ln\,(p(\mathbf{x}|\omega_i)) + ln\,(P(\omega_i)) \\ goal\ldots &\stackrel{?}{=} simple\ function\ of\ \mathbf{x} \end{aligned}$$

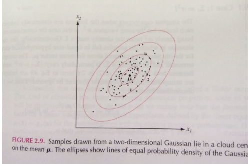
Example: Bayesian Gaussian Classifier for Equal Prior Probabilities



- $g_i(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_i)^2}, \ \mu_1 = 0, \ \mu_2 = 1$
- $g_1(x) \overset{?}{>} g_2(x)$
- $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_1)^2} \overset{?}{>} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_2)^2}$
- $e^{-\frac{1}{2}(x-\mu_1)^2} \overset{?}{>} e^{-\frac{1}{2}(x-\mu_2)^2}$
- $-\frac{1}{2}(x-\mu_1)^2 \overset{?}{>} -\frac{1}{2}(x-\mu_2)^2$
- $(x-\mu_1)^2 \overset{?}{<} (x-\mu_2)^2$
- $x \overset{?}{<} \frac{1}{2}$, much faster to execute!

Next:

Normal Probability Density Function



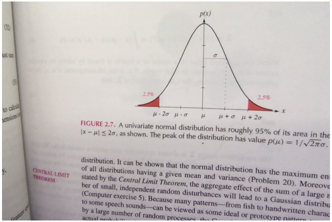
scalar Gaussian random variable is characterized by
mean: μ, variance: σ^2

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

d-dimensional Gaussian random vector is characterized by its mean vector μ and covariance matrix $\mathbf{\Sigma}$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^t \mathbf{\Sigma}^{-\mathbf{1}}(\mathbf{x}-\mu)\right]$$

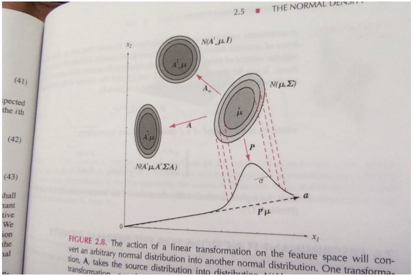
Gaussian Moments



- mean:
  $\mu = E[\mathbf{x}] = \int_{-\infty}^{\infty} p(\mathbf{x})\mathbf{x}d\mathbf{x} = [\mu_1, \mu_2, \ldots \mu_d]^t$
- covariance matrix:
  $\boldsymbol{\Sigma} = \boldsymbol{E}\left[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^{\boldsymbol{t}}\right]$
- $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^t$ (symmetric,positive semi-definite)
- Average Value of
  $f(\mathbf{x})$: $E[f(\mathbf{x})] = \int_{-\infty}^{\infty} f(\mathbf{x})p(\mathbf{x})d\mathbf{x}$
- Example: $f(\mathbf{x}) = -ln(p(\mathbf{x})) \Longrightarrow$ entropy, $H(\mathbf{x})$
- $H(x) = -E[ln(p(x))] = \frac{1}{2} + \log_2 \sqrt{2\pi\sigma^2}$
- Gaussian has largest entropy of any continuous r.v. having same mean μ and variance σ^2
- independence $\Leftrightarrow$ uncorrelatedness

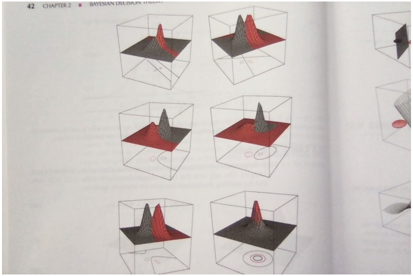
Linear Transformations of Gaussian Vectors

- $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$
- let $\mathbf{A}$ be a deterministic matrix, and let $\mathbf{y} = \mathbf{A}^t\mathbf{x}$
- $\mathbf{y} \sim \mathcal{N}(\boldsymbol{A}^t\mu,\ \boldsymbol{A}^t\mathbf{\Sigma}\boldsymbol{A})$
- special case: $\mathbf{\Sigma} = \mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^t$, where
  - columns of $\mathbf{\Phi}$ are eigenvectors of $\mathbf{\Sigma}$ (orthonormal set)
  - diagonal of $\mathbf{\Lambda}$ contain the eigenvalues of $\mathbf{\Sigma}$
  - if $\mathbf{A}_w = \mathbf{\Phi}\mathbf{\Lambda}^{-1/2}$, then $\boldsymbol{A}_w^t\mathbf{\Sigma}\boldsymbol{A}_w = \boldsymbol{I}$ (whitening transformation)



Next:

Minimum Error Probability Discriminant Functions



- Bayes Gaussian Discriminant Function:

$$g_i(\mathbf{x}) = \ln(p(\mathbf{x}|\omega_i)) + \ln P(\omega_i)$$

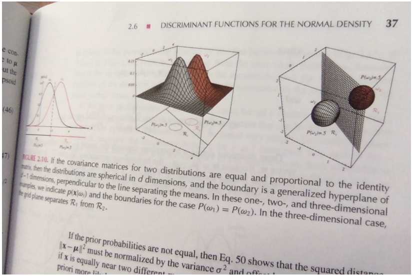
  - for each $\mathbf{x}$, choose class ω_k if
  - $k = \arg\max_i g_i(\mathbf{x})$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu_i})^t \boldsymbol{\Sigma_i^{-1}}(\mathbf{x} - \boldsymbol{\mu_i}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma_i}| + \ln P(\omega_i)$$

- $(\mathbf{x} - \boldsymbol{\mu_i})^t \boldsymbol{\Sigma_i^{-1}}(\mathbf{x} - \boldsymbol{\mu_i})$ is the *Mahalanobis* squared *distance* from $\mathbf{x}$ to $\boldsymbol{\mu_i}$
- what are the shape of the decision boundaries in feature space?

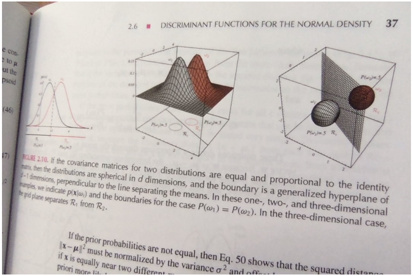
White Gaussian Vectors

- spherical Gaussian vectors: $\boldsymbol{\Sigma_i = \sigma^2 I}$
- $g_i(\mathbf{x}) = \frac{1}{2\sigma^2}\left[\mathbf{x^t x} - 2\boldsymbol{\mu_i^t x} + \boldsymbol{\mu_i^t \mu_i}\right] + \ln P[\omega_i]$
- remove terms & factors common to all i
- $g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0}$,
- $\mathbf{w}_i = \frac{1}{\sigma^2}\mu_i$, $w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu_i^t \mu_i} + \ln P[\omega_i]$



- w_{i0} is the ith *bias* or *threshold*
- $\mathbf{w}_i^t\mathbf{x}$ is a linear operator on the feature vector
- a *linear machine* is a classifier which uses such $g_i(\mathbf{x})$

Decision Boundaries for White Gaussian Vectors



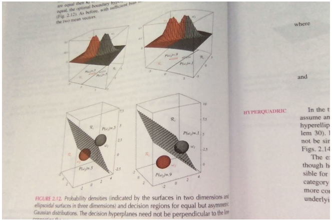
- boundary is a hyperplane
  - orthogonal to $\mathbf{w}_{12}$,
  - passes through $\mathbf{x}_0$
- decision regions are contiguous

- $g_1(\mathbf{x}) = g_2(\mathbf{x})$ specifies $\mathbf{x}$ on a decision boundary, or
- $\mathbf{w}_{12}^t(\mathbf{x} - \mathbf{x}_0) = 0$, with
- $\mathbf{w}_{12} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$, and
- $\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) - \frac{\sigma^2}{\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2} \ln \frac{P[\omega_1]}{P[\omega_2]}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$

Commonly-colored Gaussian Vectors

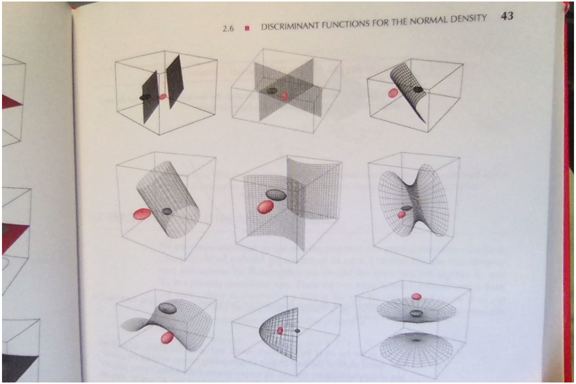
$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu_i}, \boldsymbol{\Sigma})$$

- $\boldsymbol{\Sigma_i} = \boldsymbol{\Sigma}$
- $\mathbf{y} = \mathbf{A_w x}$, then colored Gaussian becomes white in y-feature space
- in x-feature space, the discriminant functions are:
- $g_i(\mathbf{x}) = -\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu_i})^t \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu_i}) + \ln P[\omega_i]$ is equivalent to:
- $g_i(\mathbf{x}) = \mathbf{q_i^t x} + c_{i0}$, where
  - $\mathbf{q}_i = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_i}$
  - $c_{i0} = -\frac{1}{2}\boldsymbol{\mu_i^t}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu_i} + \ln P[\omega_i]$



- boundary is a hyperplane
- not orthogonal to mean difference

Decision Boundaries - Arbitrary Gaussian Vectors



$$w_{i0} = -\frac{1}{2}\mu_i^t \boldsymbol{\Sigma}_i^{-1}\mu_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P[\omega_i]$$

- $g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0}$
- $\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}$
- $\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\mu_i$

Binary Error Probability Calculation

- Bayes Formula:
- $P[\omega_j|\mathbf{x}] = p(\mathbf{x}|\omega_j)P[\omega_j]/p(\mathbf{x})$
- $p(\mathbf{x}) = \sum_{j=1}^{2} p(\mathbf{x}|\omega_j)P[\omega_j]$
- Bayes Decision Rule (uniform costs):
- decide ω_1 if $P[\omega_1|\mathbf{x}] > P[\omega_2|\mathbf{x}]$

- $P(error|\mathbf{x}) = min(P(\omega_1|\mathbf{x}), P(\omega_2|\mathbf{x}))$
- $P(error) = \int P(error|\mathbf{x})d\mathbf{x}$ p(x)
- $min[a,b] \leq a^{\beta} b^{1-\beta},\ for\ a,b \geq 0,\ 0 \leq \beta \leq 1$
- Chernoff Bound
- $P(error) \leq P^{\beta}(\omega_1)P^{1-\beta}(\omega_2)\int p^{\beta}(\mathbf{x}|\omega_1)p^{1-\beta}(\mathbf{x}|\omega_2)$
- Bhattacharyya Bound: set $\beta = 1/2$
- No integration over decision regions!

Gaussian Chernoff Bounds

- $P(error) \leq P^{\beta}(\omega_1)P^{1-\beta}(\omega_2)e^{-k(\beta)}$,

$k(\beta) =$
$$\frac{\beta(1-\beta)}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})^t\left[(1-\beta)\boldsymbol{\Sigma_1} + \beta\boldsymbol{\Sigma_2}\right]^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) + \frac{1}{2}\ln\frac{|(1-\beta)\boldsymbol{\Sigma_1} + \beta\boldsymbol{\Sigma_2}|}{|\boldsymbol{\Sigma_1}|^{1-\beta}|\boldsymbol{\Sigma_2}|^{\beta}}$$

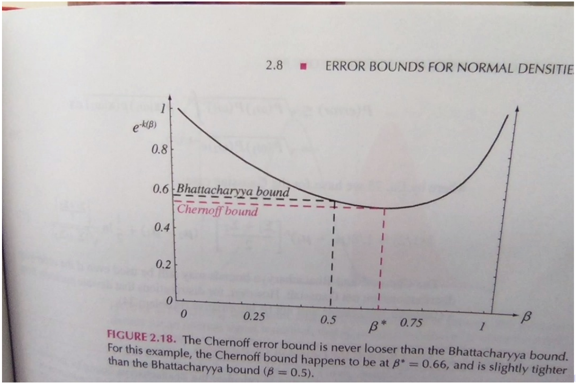


FIGURE 2.18. The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^* = 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$).

- 1 parameter to optimize!

Next:

Signal Detection Theory

- two classes: ω_1 =null class, ω_2 = alternative class
- observe $\mathbf{x}$, decide for ω_1 (negative)
  - or decide for ω_2 (positive)

Four possible events:

- *Hit:* true positive
- *False alarm:* false positive
- *Miss:* false negative
- *Correct Rejection:* true negative

Decision Events

		Condition (as determined by "Gold Standard" or "Ground Truth")		
		Positive	***Negative***	
Test outcome	***Positive***	**True Positive** T_p	**False Positive** (Type I Error) F_p	→Positive Predicted Value
	Negative	**False Negative** (Type II Error) F_n	**True Negative** T_n	→Negative Predicted Value
		↓ **Sensitivity**	↓ **Specificity**	

$= \frac{T_p}{T_p + F_n}$ $= \frac{T_n}{F_p + T_n}$

Alternative Event Labels

Medical Diagnoses:

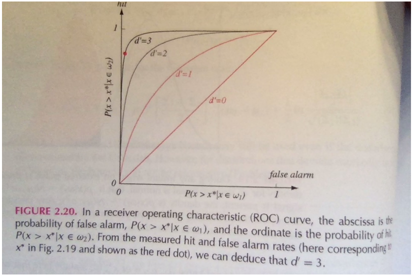
- *Sensitivity:* $1 - P[error|\omega_2] = T_p/(T_p + F_n)$
- *Specificity:* $1 - P[error|\omega_1] = T_n/(T_n + F_p)$

Information Retrieval:

- *Precision:* $T_p/(T_p + F_n)$ p
- *Recall:* $T_p/(T_p + F_n) = ?$

		correct result /	classification
		C1	C2
obtained result / classification	C1	**tp** (true positive)	**fp** (false positive)
	C2	**fn** (false negative)	**tn** (true negative)

Receiver Operating Characteristic Curves



- ROCs display Type I error rate (false alarm prob) vs...
- 1-Type II error rate (hit prob)
- shown for scalar Gaussians, common variance
- discriminability $d' = |\mu_1 - \mu_2| / \sigma$

2-Column Template