Probability, Bayes Decision Theory

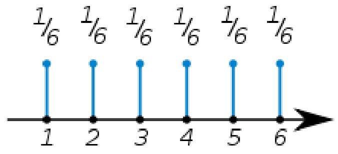
by Ahmet Sacan

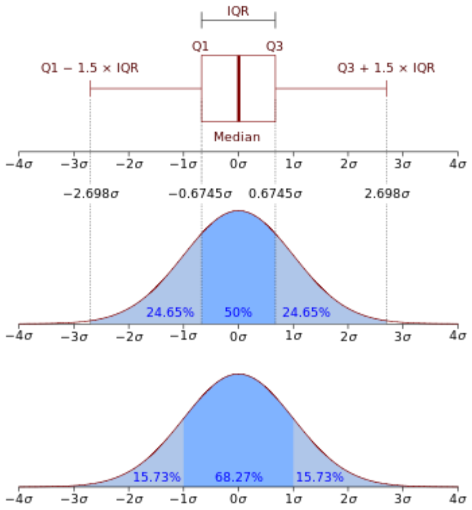
Probability

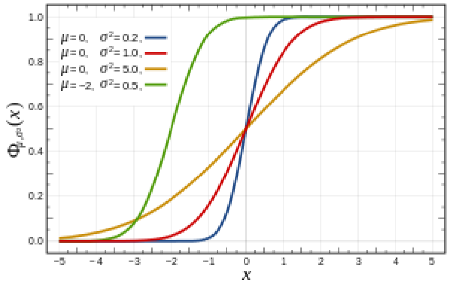
- "Probability is a way of expressing knowledge or belief that an event will occur or has occurred." (wikipedia)

Probability

- Probability mass function
  - a function that gives the probability that a <u>discrete random variable</u> is <u>exactly equal</u> to some value
- Probability density function
  - a function of a <u>continuous random variable</u>, whose integral across an interval gives the probability that the value of the variable lies within the same interval.
- Cumulative distribution function
  - cumulative distribution function (CDF) of a real-valued random variable V evaluated at x, is the probability that V will take a value less than or equal to x.
    - cdf(x)=$cdf(x) = \int_{-inf}^{x} pdf(u)du$







Definitions and images are from wikipedia

Probability

- Joint probability
  - P(A&B), P(A,B), $P(A \cap B)$
- Conditional probability
  - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - $P(A \cap B) = P(A|B) * P(B)$
- Independent variables
  - $P(A|B) = P(A)$
  - $P(A \cap B) = P(A) * P(B)$
- Bayes Rule
  - $P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$

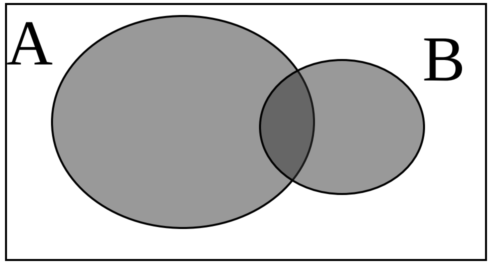
"Probability of a random day <u>A</u>: being Saturday <u>B</u>: given that it is a weekend."

"Probability of a 2 being rolled on die roll, given a coin toss resulted in a tail."

"Probability of a 2 being rolled on die roll, AND a coin toss resulting in a tail."

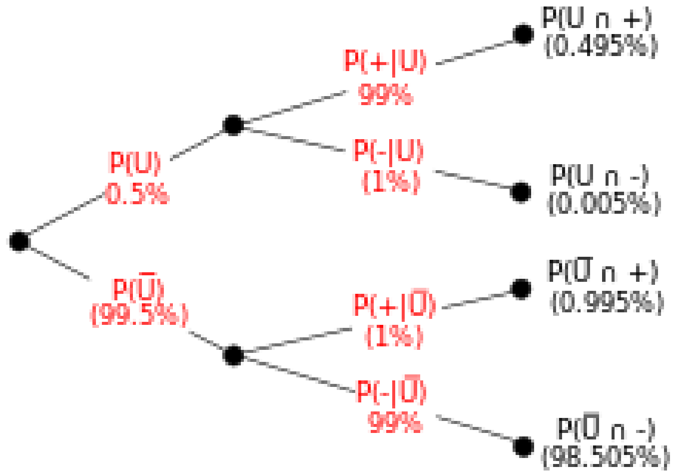
Conditional Probability

- Deals with partial information
- Example: Let
  - $P(A)=\{1,2,3,4\}$, $P(B)=\{1,2,5\}$ in a die throw
- What is the probability of A, given that we know B has occurred.
  - $P(A|B)=P(A,B)/P(B)$



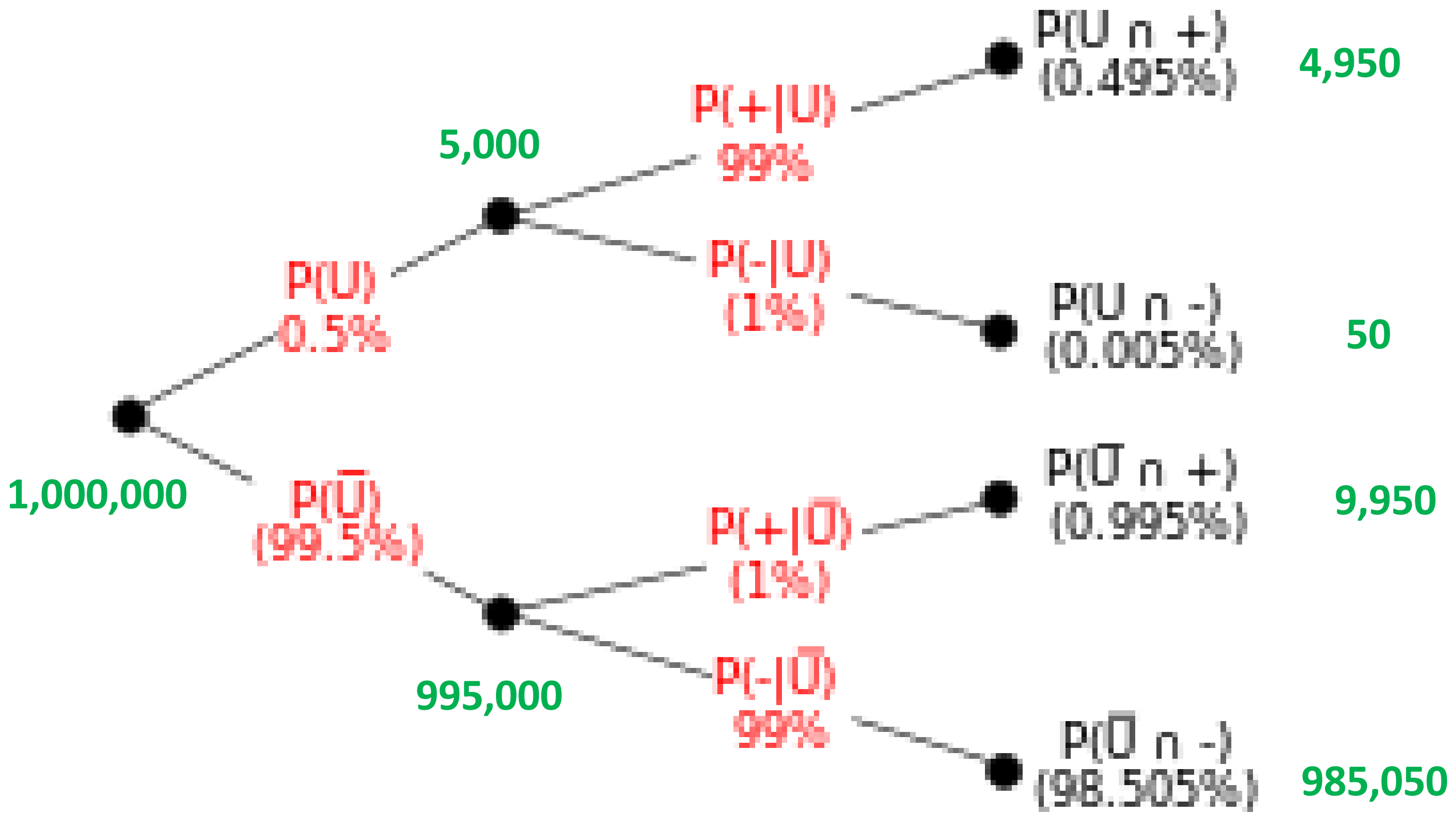
Bayes' Rule

- Suppose a drug test is 99% sensitive and 99% specific. That is, the test will produce 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose that 0.5% of people are users of the drug. If a randomly selected individual tests positive, what is the probability that he is a user?



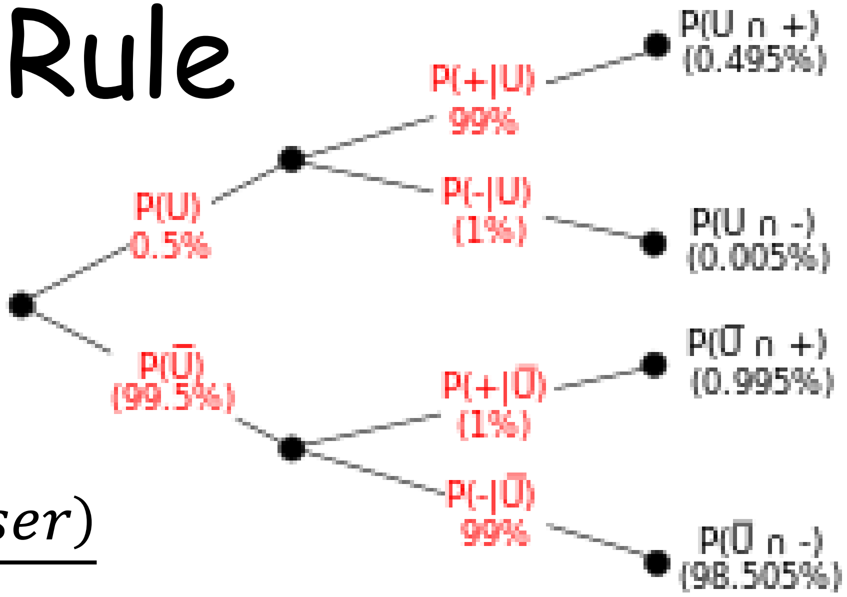
https://en.wikipedia.org/wiki/Bayes'_theorem

Bayes' Rule



$$P(User|+) = \frac{4,950}{4,950 + 9,950} = \frac{4,950}{14900} = 0.33 = 33\%$$

Bayes' Rule



$$P(User|+) = \frac{P(+|User)P(User)}{P(+)}$$

$$= \frac{P(+|User)P(User)}{P(+,User) + P(+|NonUser)}$$

$$= \frac{P(+|User)P(User)}{P(+|User)P(User) + P(+|NonUser)P(NonUser)}$$

$$= \frac{0.99*0.005}{0.99*0.005 + 0.01*0.995}$$

$$\approx 33\%$$

Probability

$$p(x) = \sum_i p(x, y_i) = \sum_i p(x \mid y_i) p(y_i)$$

$$P(x) = \int P(x, y) dy = \int P(x \mid y) p(y) dy$$

- P(run) = P(run,rain) + P(run,sunny) + P(run,snow)

  = $\sum_i P(run, weather_i)$

  = P(run|rain) P(rain) + P(run|sun)P(sun) + P(run|snow)P(snow)

  = $\sum_i P(run|weather_i) P(w_i)$

Bayes Decision Theory

- Example: identify fresh vs. rotten walnut, given its weight.

https://sevalorhan.wordpress.com/tag/ceviz-ici-ask/

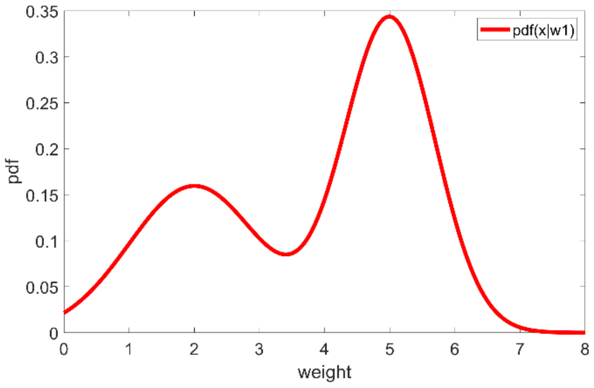
https://depositphotos.com/31321921/stock-photo-rotten-nut-on-white.html

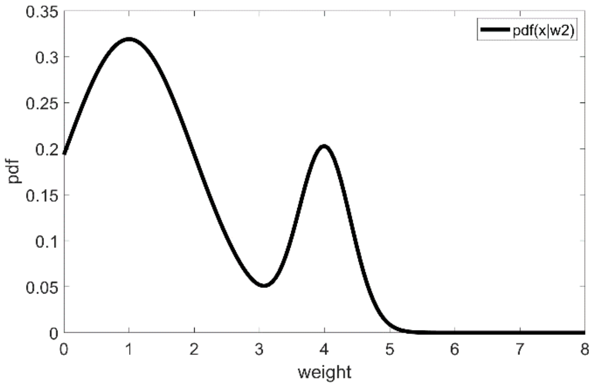
Bayes Decision Theory

- Example: identify fresh vs. rotten walnut, given its weight.

https://sevalorhan.wordpress.com/tag/ceviz-ici-ask/

https://depositphotos.com/31321921/stock-photo-rotten-nut-on-white.html



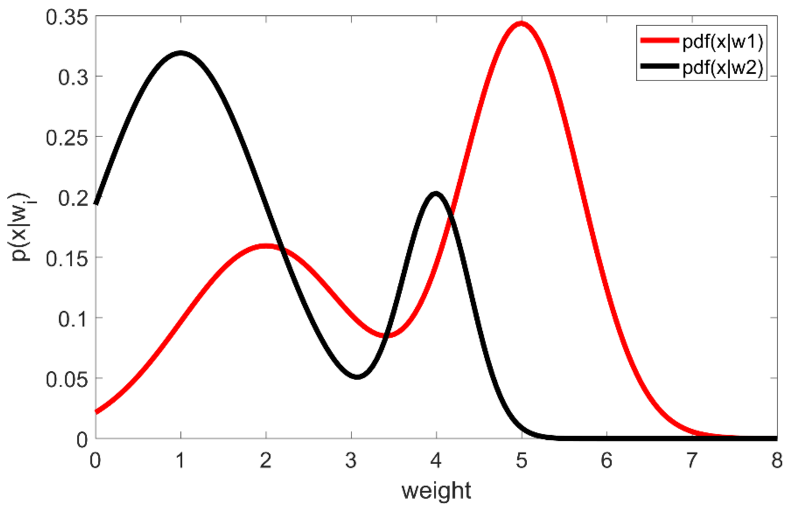


Bayes Decision Theory

- Example: identify fresh vs. rotten walnut, given its weight.

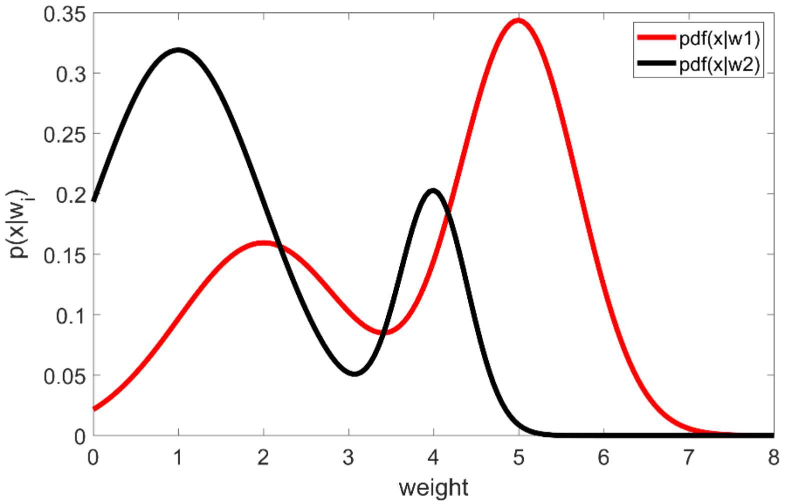
https://sevalorhan.wordpress.com/tag/ceviz-ici-ask/

https://depositphotos.com/31321921/stock-photo-rotten-nut-on-white.html



Bayes Decision Theory

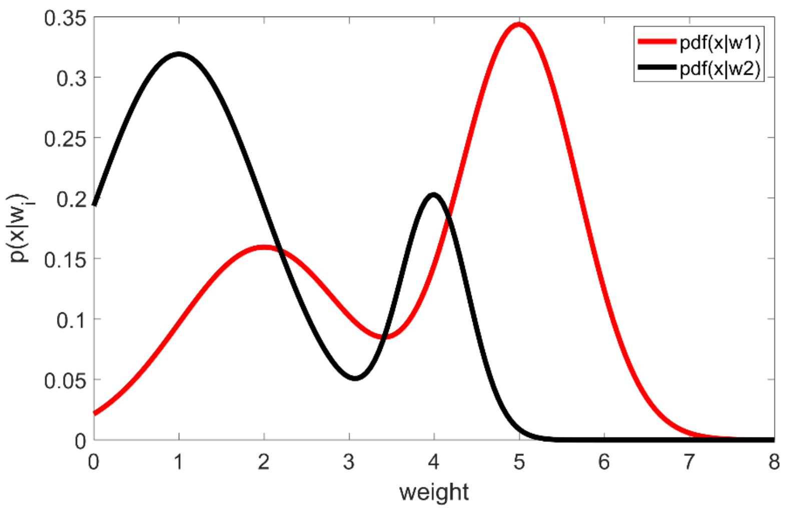
- Classification task:
  - w1/w2: fresh vs. rotten walnut
  - x: weight
- *Apriori* probability
  - background frequency of fresh, frequency of rotten
- Class-conditional probability density function: p(x|w)
- $P(w_1|x) = p(x|w_1) * P(w_1)/p(x)$
- $P(w_2|x) = p(x|w_2) * P(w_2)/p(x)$
- $p(x) = \sum p(x|w_i) * P(w_i)$

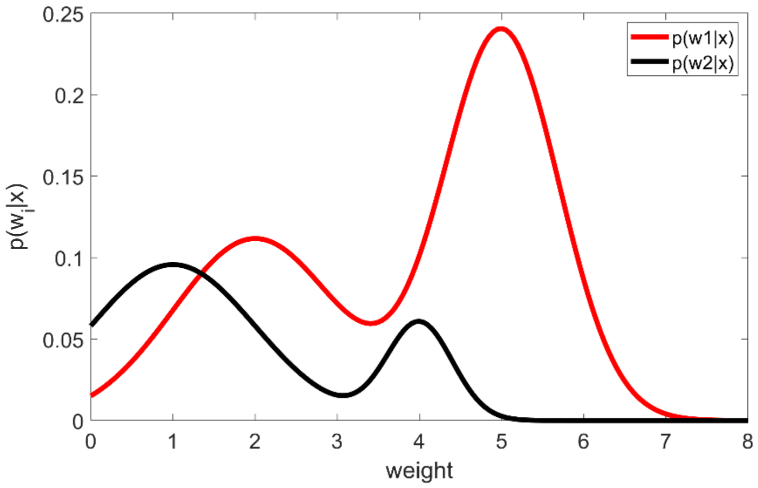


Bayes Decision Theory

$$P(w|x) = \frac{p(x|w) * P(w)}{p(x)}$$

$$posterior = \frac{likelihood * prior}{evidence}$$





- Class conditional probability (aka "likelihood") is not enough.
- Select the class with highest posterior probability.

Bayes Decision Rule

- evidence, p(x), is just a scaling factor (to ensure P(w1|x)+P(w2|x)=1)
- Bayes Decision Rule
  - Decide w1 if: p(x|w1)*P(w1) > p(x|w2)*P(w2)
  - Otherwise, decide w2
- Likelihood ratio:

$$\frac{P(w1|x)}{P(w2|x)} = \frac{p(x|w1)P(w1)}{p(x|w2)P(w2)}$$

  - Decide w1 if likelihood ratio >1
  - Otherwise, decide w2.

Probability of Error

- P(error|x) =
  - P(w1|x) if we decide w2
  - P(w2|x) if we decide w1

$$P(error) = \int P(error \mid x) p(x) dx$$

- P(error|x) = min[P(w1|x),P(w2|x)]

Loss/Cost Function

- Feature space
- Loss function L(ai|wj)
  - Measures how costly each decision is.
  - Cost of action ai when the truth was wj
- Risk of action
  - $R(ai|x) = Sum_j(L(ai|wj)*P(wj|x))$
  - Expected loss
  - Conditional risk
- Decision Rule: a function a(x) that minimizes the overall risk.
  - Calculate risk for each action, select the minimum Bayes Risk

https://cooking.stackexchange.com

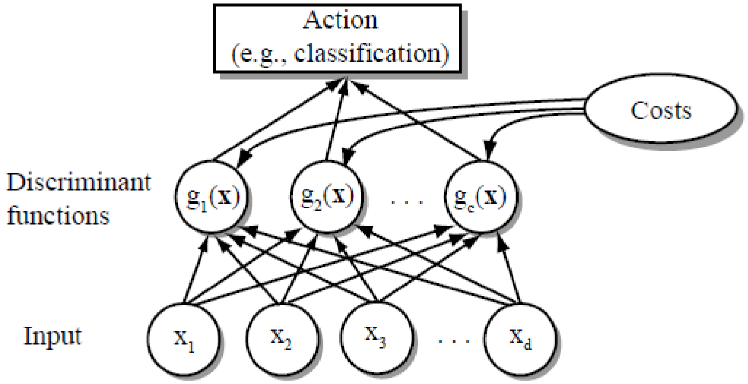
Multi-category case

- Discriminant function

$$g_i(x) = P(w_i \mid x) = \frac{p(x \mid w_i)P(w_i)}{\sum_{j=1}^{c} p(x \mid w_j)P(w_j)}$$

$$g_i(x) = p(x \mid w_i)P(w_i)$$

$$g_i(x) = \ln p(x \mid w_i) + \ln P(w_i)$$



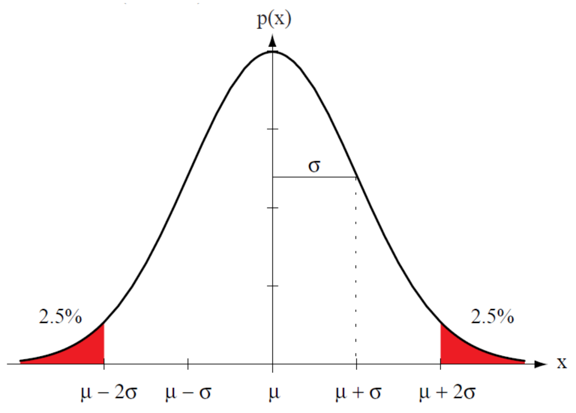
Dichotomizer (two-category case)

- $g(x) = g_1(x) - g_2(x)$

- $g(x) = \ln p(x|w_1) + \ln P(w_1) - (\ln p(x|w_2) + \ln P(w_2))$

- $g(x) = \ln \frac{p(x|w_1)}{p(x|w_2)} + \ln \frac{P(w_1)}{P(w_2)}$

Univariate Normal Density



- Expected value: $E[X] = \sum x_i p_i$
- Mean
- Variance: $Var(X) = E[(X - \mu)^2]$
- Entropy: $-\sum P(x_i) \log_2(P(x_i))$
  - Fundamental uncertainty in the values of randomly selected points from a distribution
- Normal density has the maximum entropy of all distributions having a given mean and variance.
- Gaussian is good for
  - an ideal prototype pattern, corrupted by large number of random processes.

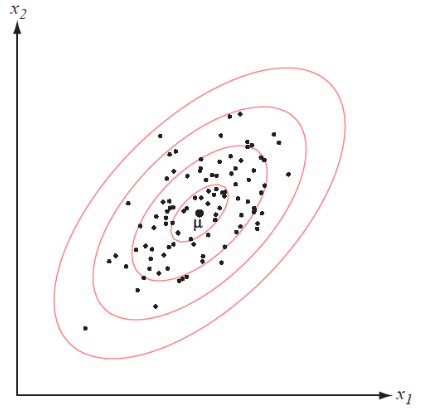
Discriminant Function for Univariate Normal Density

- $P(x|w_i) = normpdf(x|w_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $P(w|x) = \frac{P(x|w)P(w)}{P(x)} \approx P(x|w)P(w)$
- Take the logarithm (for convenience) to define the discriminant function:

$$g_i(x) \approx -\frac{(x-\mu)^2}{2\sigma^2} - \ln(\sigma) + \ln P(w_i)$$

Above, we are too lazy to write σ_i and μ_i.

Multivariate Normal Density



- Covariance matrix is always
  - Symmetric
  - Semidefinite
  - →determinant is strictly positive
  - Diagonals are variances of respective xi
  - If xi and xj are statistically independent, then their covariance is zero.
  - If all non-diagonal entries are zero, p(x) equals the product of the univariate densities of components of x.

Case 1: IID features

- Statistically independent features, each with variance σ^2, i.e., $\Sigma = \sigma^2 I$.

$$
\begin{aligned}
\mathrm{g}_i(x) &= -\frac{(x-\mu)^t \Sigma^{-1} (x-\mu)}{2} - \ln\left(\sqrt{|\Sigma|}\right) + \ln P(w_i) \\
&= -\frac{\|x-\mu\|^2}{2\sigma^2} - \ln(\sigma) + \ln P(w_i) \\
&= -\frac{x^t x - 2\mu^t x + \mu^t \mu}{2\sigma^2} - \ln(\sigma) + \ln P(w_i) \\
&= -\frac{x^t x}{2\sigma^2} + \frac{\mu^t}{2\sigma^2} x - \frac{\mu^t \mu}{2\sigma^2} - \ln(\sigma) + \ln P(w_i)
\end{aligned}
$$

Case 1: IID features & Same σ (Identical Class Distributions)

- If we also assume all classes have the same σ, we can drop the terms that have the same values for all classes:

$$g_i(x) = \frac{\mu^t}{2\sigma^2} x - \frac{\mu^t \mu}{2\sigma^2} + \ln P(w_i)$$

- Which is a linear discriminant function ("linear machine")

$$g_i(x) = A_i x - A_{i0}$$

where $\mathrm{A_i} = \frac{\mu^t}{2\sigma^2}$ and $A_{\mathrm{i0}} = \frac{\mu^t \mu}{2\sigma^2} + \ln P(w_i)$

Case 1: IID, Same σ

- Decision surfaces for a linear machine are pieces of hyperplanes defined by the linear equations:

$$g_i(x) = g_j(x)$$

- Observations:
  - Equations define a hyperplane through A_{i0} and orthogonal to the vector A_i.
  - If variance is small relative to the squared distance $\left\|\mu_i - \mu_j\right\|^2$, the position of the boundary is relatively insensitive to the exact values of the prior probabilities. ("Minimum distance classifier", "template matching" against ideal prototype)

Case 1: IID, Same σ, Two classes

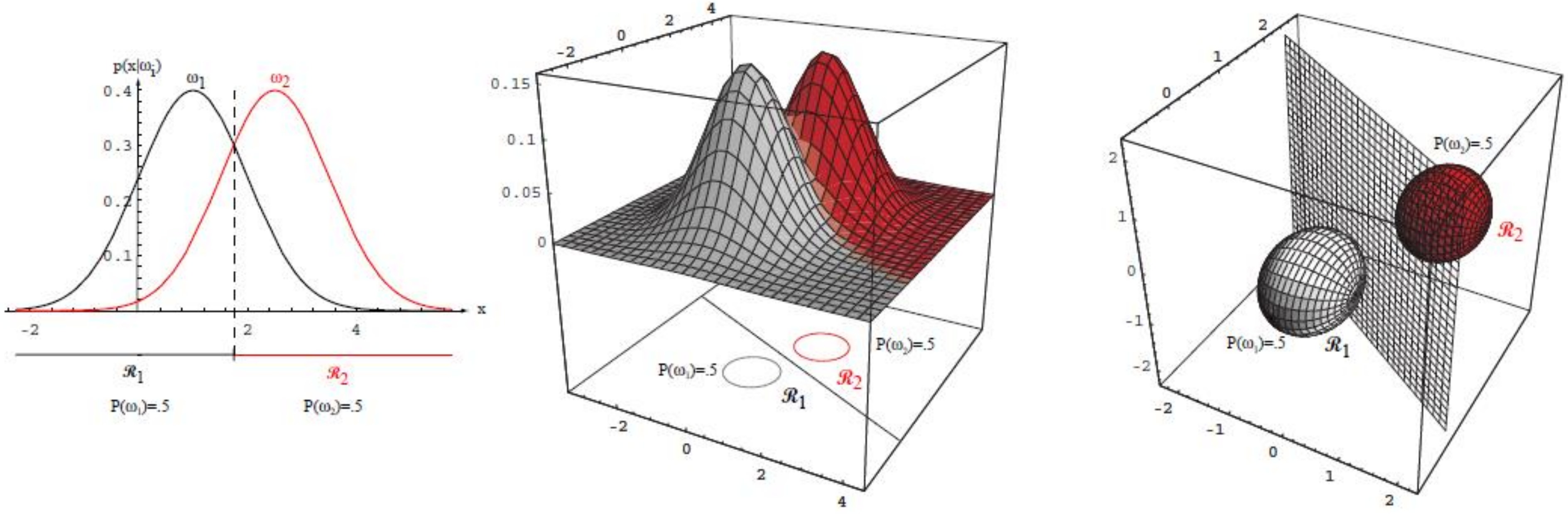
- Decision boundary can be obtained by solving for:

$$g_1(x) - g_2(x) = 0$$

$$\frac{(\mu_1 - \mu_2)^t}{2\sigma^2} x - \frac{\mu_1^t \mu_1 - \mu_2^t \mu_2}{2\sigma^2} + \ln \frac{P(w_1)}{P(w_2)} = 0$$

$$x = (\mu_1 - \mu_2) \frac{2\sigma^2 \ln \frac{P(w_2)}{P(w_1)} + \|\mu_1\|^2 - \|\mu_2\|^2}{\left\|\mu_i - \mu_j\right\|^2}$$

Case 1: IID, Same σ, Two classes



- Decision boundary is a generalized hyperplane of d-1 dimensions, perpendicular to the line separating the means.

Case 2: identical covariance

- $g_i(x)=\ldots$
- Again, linear

Case 3: arbitrary covariances

- Quadratic

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0}, \tag{64}$$

where

$$\mathbf{W}_i = -\frac{1}{2}\boldsymbol{\Sigma}_i^{-1}, \tag{65}$$

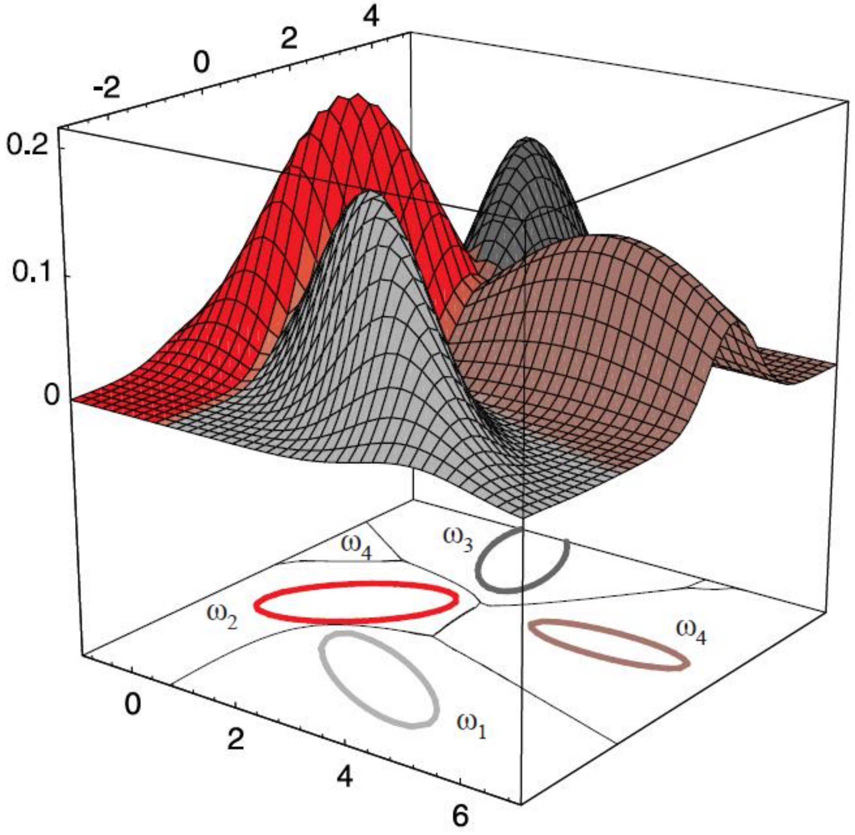
$$\mathbf{w}_i = \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i \tag{66}$$

and

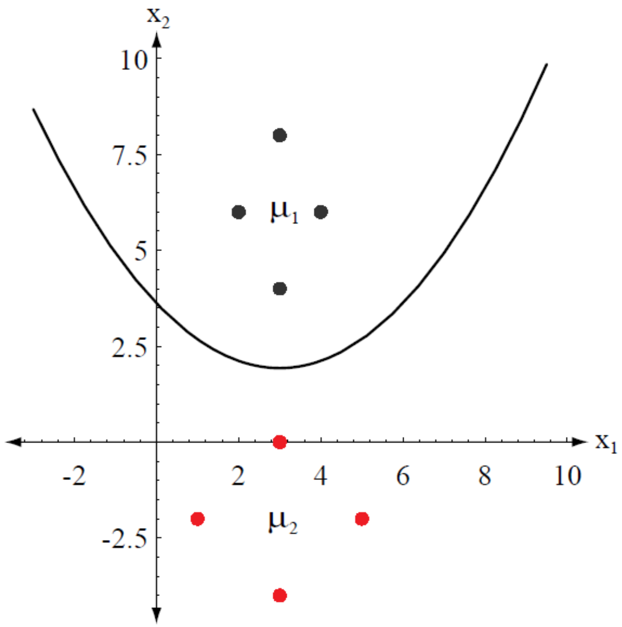
$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i). \tag{67}$$

Arbitrary Covariances, Multiple classes

- Decision boundary can be complex
- Example: 4 normally distributed classes in 2-D.



Example: 2D, independent features, arbitrary covariances



- Assume P(w1) = P(w2) = 0.5

$$\mu_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \Sigma_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix}$$

$$\mu_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \Sigma_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

- Plug in to the general form of discriminant functions, and solve for $g_1(x) = g_2(x)$
- $x_2 = 0.19x_1^2 - 1.1x_1 + 3.5$

Summary

- $P(w_1|x) = \frac{p(x|w_1)P(w_1)}{p(x)}$

- $p(x) = \sum p(x|w_i) * P(w_i)$