Continuous Random Variable

- Recall that a discrete RV has a countable range
- Some RVs, however, can take on any value in an interval
  - Examples:
    - X = Waiting time for a shuttle bus
    - W = Water level at a reservoir
    - R = Voltage across a resistor
- Suppose that we drop a ball on a unit interval [0, 1]. What is the probability that it will land in the middle (i.e., 0.5)?

Unlike with discrete RVs, we cannot identify a probability for each possible value in the range

Instead, we talk about the probability that the RV falls in some "range"

Continuous Random Variable

Definition: A random variable (RV) X is said to be a "continuous" RV if its (cumulative) distribution function can be written as

$$F_X(x) := \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f(u)\, du$$

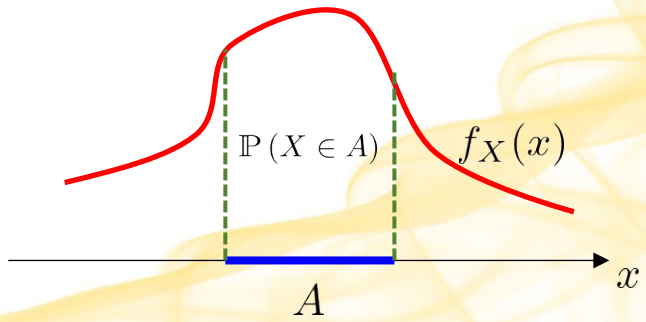
for some integrable function $f_X : \mathbb{R} \to [0, \infty)$

Definition: The function $f_X : \mathbb{R} \to [0, \infty)$ is called the (probability) density function (PDF)

Use of PDF:

Given any $A \subset \mathbb{R}$,

$$\mathbb{P}(X \in A) = \int_A f_X(x)\, dx$$



Properties of PDFs

- Before we state the properties, I emphasize that $f_X(x) \neq \mathbb{P}(X = x)$

- Properties

  (i) For $x_1, x_2 \in \mathbb{R}, x_1 \leq x_2$, $\mathbb{P}(x_1 < X \leq x_2) = \int_{x_1}^{x_2} f_X(u)\, du$

  (ii) $\int_{\mathbf{R}} f_X(x)\, dx = 1$

  (iii) For all $B \in \mathcal{B}$ (Borel σ-field), $\mathbb{P}(X \in B) = \int_B f(u)\, du$

Expectation

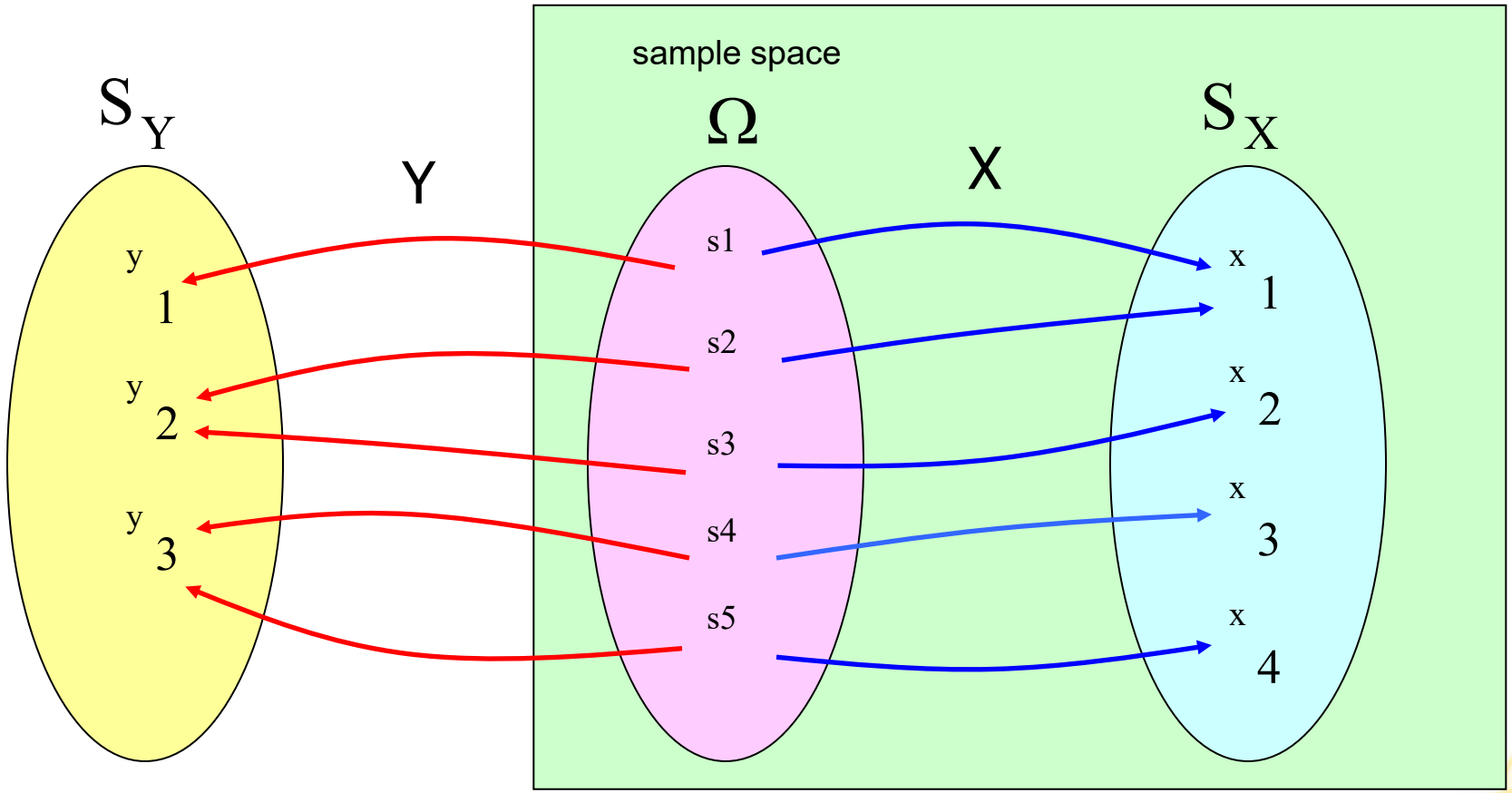
- For a discrete RV Y with PMF $p_Y : \mathbb{R} \to [0, 1]$, the expectation or expected value of Y is given by

$$\mathbb{E}[Y] = \sum_{y \in S_Y} y \cdot p_Y(y)$$

- For a continuous RV X with a PDF $f_X : \mathbb{R} \to \mathbb{R}_+$, its expectation or expected value is given by

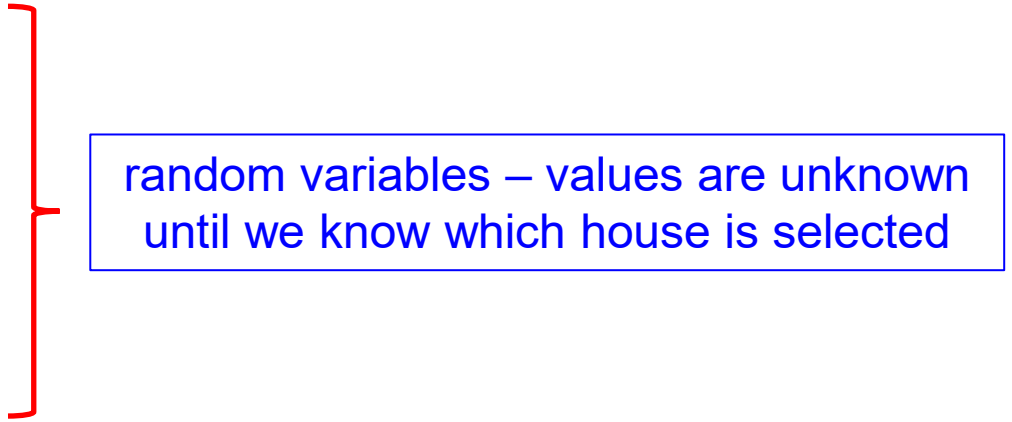
$$\mathbb{E}[X] = \int_{\mathbb{R}} x \cdot f_X(x)\ dx$$

Multiple RVs



Multiple RVs

- Example: Randomly choose a house in College Park
  - R = # of rooms
  - B = # of bathrooms
  - G = # of garage parking spaces
  - L = lot size (square feet)
  - A = age of the house (years)



- In machine learning (ML), feature values often modeled as random variables with some unknown distribution which may depend on class or label

Multiple Discrete RVs

- Joint Probability Mass Function (PMF) of two discrete RVs: $p_{X,Y} : \mathbb{R}^2 \to [0,1]$, where

$$p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x,\ Y(\omega) = y\})$$

  - Recall: for a single RV X, $p_X(x) = \mathbb{P}(X = x), \quad x \in \mathbb{R}$

- Range of (X,Y): set of possible values of the pair (X,Y)

$$S_{X,Y} = \{(x,y) \mid p_{X,Y}(x,y) > 0\}$$

Multiple Discrete RVs

- Example: Consider a randomly selected house in a neighborhood

$$X = \text{number of bedrooms}$$

$$Y = \text{number of bathrooms}$$

$$p_{X,Y}(x,y) = \mathbb{P}\left(X = x, Y = y\right), \;\; x, y \in \{0, 1, \ldots\}$$

$$p_{X,Y}(4,2) = \mathbb{P}\left(4 \text{ bedrooms and } 2 \text{ bathrooms}\right)$$

Marginal PMF

- Suppose that we are given the joint PMF, but we are only interested in one of the two RVs

$$p_X(x) = \sum_{y \in S_Y} p_{X,Y}(x,y) \quad \text{and} \quad p_Y(y) = \sum_{x \in S_X} p_{X,Y}(x,y)$$

- Example: Consider a randomly selected house in a neighborhood

$$p_X(4) = \mathbb{P}(X = 4) = \sum_{y=0}^{\infty} p_{X,Y}(4,y)$$

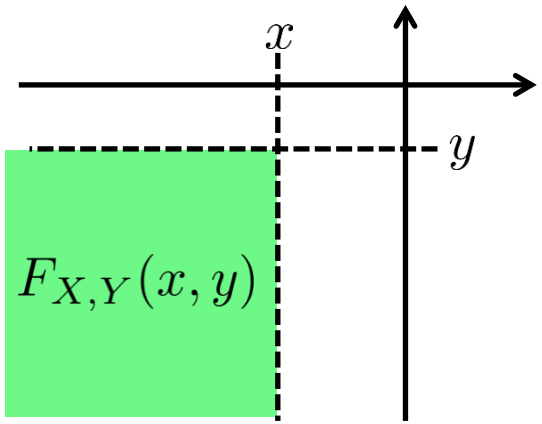
- Probability a randomly selected house has 4 bedrooms

Multiple Continuous RVs

- Joint Probability Density Function (PDF) of two continuous RVs: $f_{X,Y} : \mathbb{R}^2 \to [0, \infty)$, where

$$F_{X,Y}(x, y) := \mathbb{P}(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v)\, dv\, du$$

for all $x, y \in \mathbb{R}$



Given $A \in \mathbb{R}^2$,

$$\mathbb{P}((X, Y) \in A) = \int \int_{A} f_{X,Y}(x, y)\, dy\, dx$$

- Range of (X,Y): set of possible values of the pair (X,Y)

$$S_{X,Y} = \{(x, y) \mid f_{X,Y}(x, y) > 0\}$$

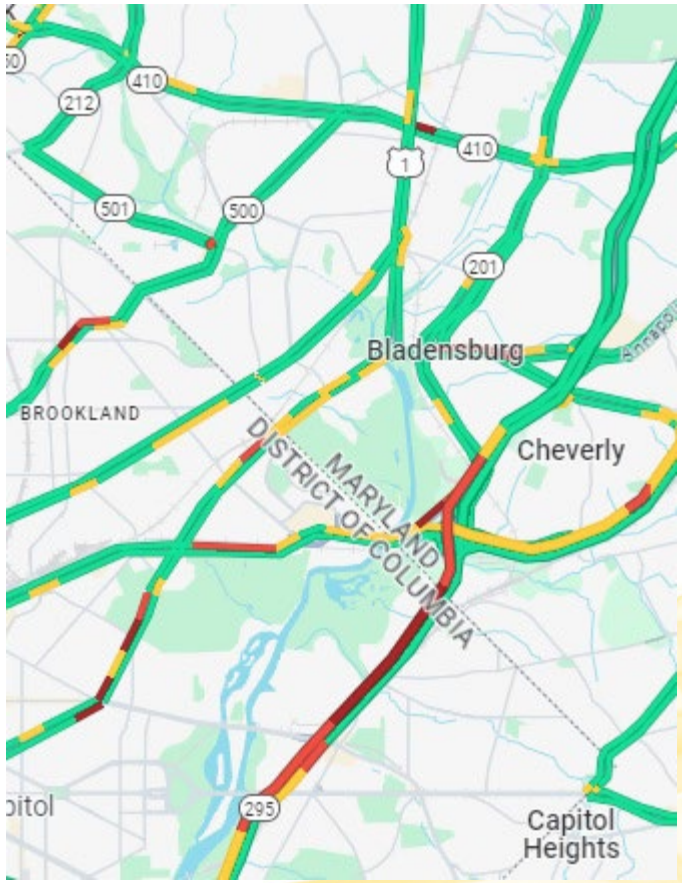
Multiple Continuous RVs

- Example: Traffic volume on different roads at 7:30 AM

$$X = \text{traffic on Route 1(going south)}$$
$$Y = \text{traffic on 295 (going south))}$$

$$\mathbb{P}\left(X \in [x_1, x_2], Y \in [y_1, y_2]\right) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{X,Y}(x, y)\, dy\, dx$$

$$x_1 < x_2 \;\&\; y_1 < y_2$$



Marginal PDF

- Suppose that we are given the joint PDF of X and Y, but we are only interested only in one of the two RVs

$$f_X(x) = \int_{\mathbf{R}} f_{X,Y}(x,y)\, dy \quad \text{and} \quad f_Y(y) = \int_{\mathbf{R}} f_{X,Y}(x,y)\, dx$$

- Example: Traffic volume on different roads at 7:30 AM

$$X = \text{traffic on Route 1(going south)}$$
$$Y = \text{traffic on 295 (going south))}$$

$$f_X(2.5) = \int_{\mathbb{R}} f_{X,Y}(2.5,y)\, dy$$

Independence

- Suppose that X and Y are two RVs
- Definition: Two discrete RVs X and Y are independent if the events {X = x} and {Y = y} are independent for all x and y in $\mathbb{R}$, i.e.,

$$p_{X,Y}(x,y) = p_X(x) \times p_Y(y) \quad \text{for all } x, y \in \mathbb{R}$$

- Definition: Two continuous RVs X and Y are independent if for all x and y in $\mathbb{R}$

$$f_{X,Y}(x,y) = f_X(x) \times f_Y(y)$$

- Interpretation: Value of one RV does not affect the probability distribution of the other RV

Independence

- Example: Roll two six-sided dice

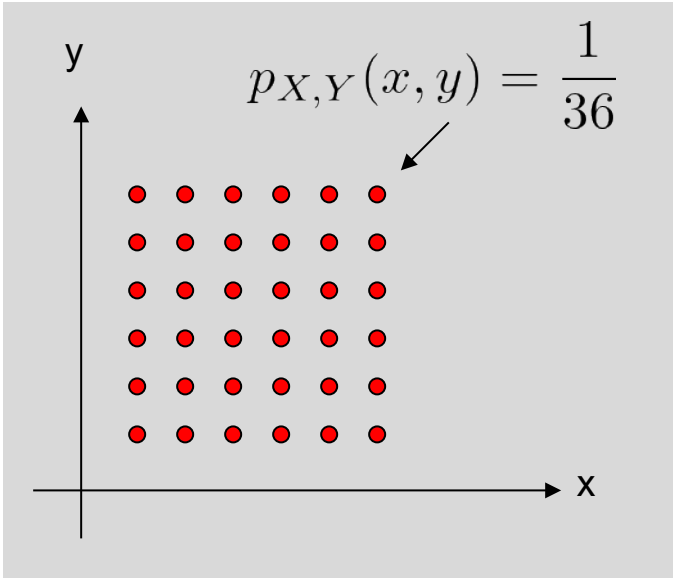
  X = number of dots on die #1

  Y = number of dots on die #2

$$p_X(x) = \begin{cases} 1/6 & x \in \{1, 2, \ldots, 6\} \\ 0 & \text{otherwise} \end{cases}$$

$$p_Y(y) = \begin{cases} 1/6 & y \in \{1, 2, \ldots, 6\} \\ 0 & \text{otherwise} \end{cases}$$

$$p_{X,Y}(x, y) = p_X(x) \times p_Y(y)$$



➡ X and Y are independent

Independence

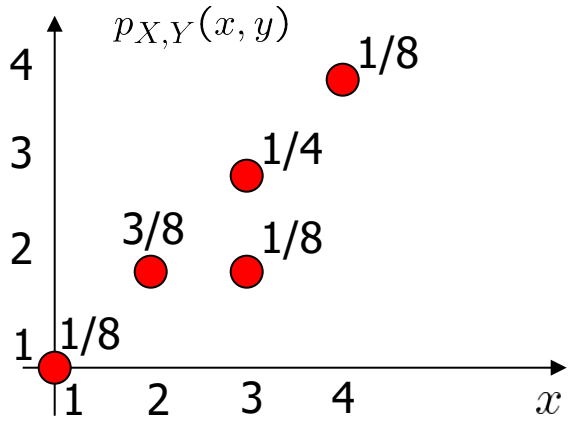
- Example #2: # of bedrooms and # of bathrooms in a randomly chosen house

X = number of bedrooms

Y = number of bathrooms

$$p_{X,Y}(2,2) = \frac{1}{4}$$

$$\neq p_X(2) \cdot p_Y(2) = \frac{3}{8} \cdot \frac{1}{4} = \frac{3}{32}$$

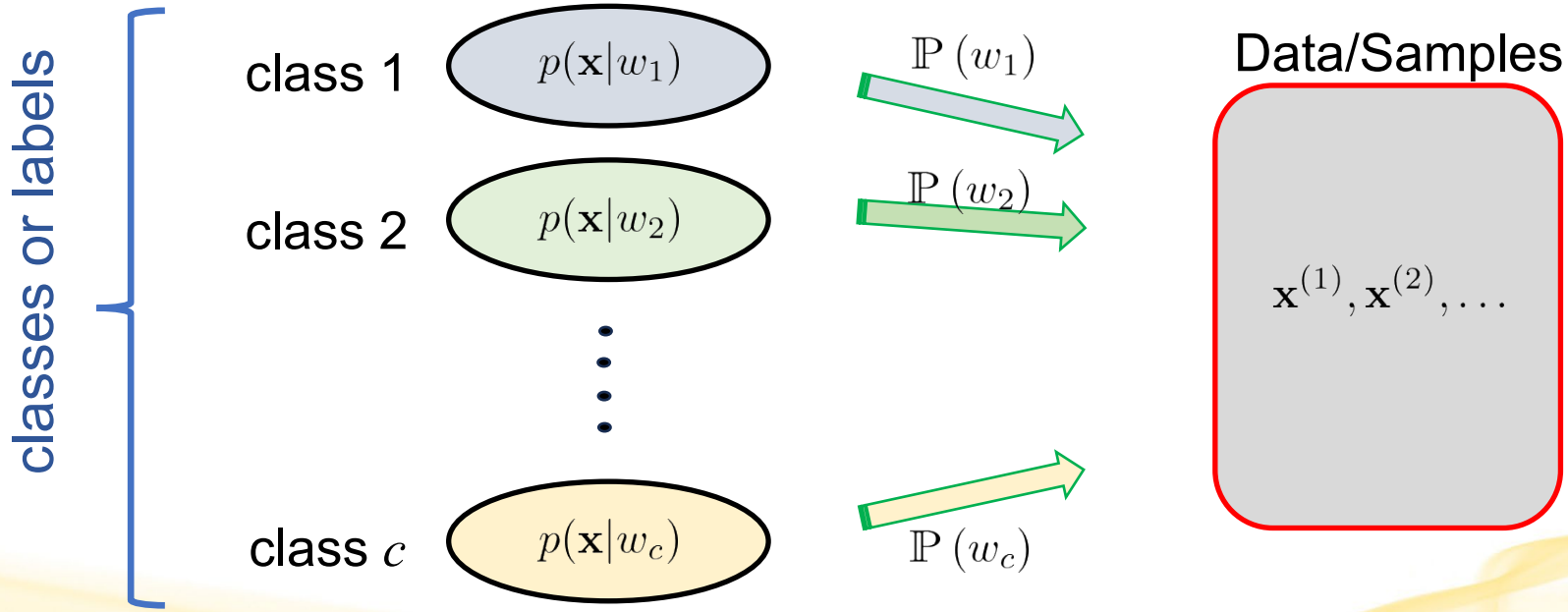


➡ X and Y not independent

Question: How could I have guessed this without doing any algebra?

Conditional Distribution

- In many cases, we will deal with situation where the distribution of observed random variables depends on some other (hidden) state
  - Example: The feature values are observed data/samples and their joint distribution of features depends on class or label



Conditional PMF

- When the observed RVs (e.g., feature values) are discrete and the (hidden) state is $\{W = w\}$, we use conditional probability mass function

$$p_w(\mathbf{x}) = \mathbb{P}\left(\mathbf{X} = \mathbf{x} | W = w\right) = \mathbb{P}\left((X_1, \ldots, X_d) = (x_1, \ldots, x_d) | W = w\right)$$

- Example: A house could be (i) single family home (S), (ii) town home (T) or (iii) condo (C). The distribution of feature values depends on the type of the house
  - R = # of bedrooms
  - B = # of bathrooms

features

$$p_C(2,2) = \mathbb{P}\left(R = 3, B = 2 | \text{condo}\right), \; p_T(4,3) = \mathbb{P}\left(R = 4, B = 3 | \text{town home}\right)$$

Conditional PDF

- When the observed RVs (e.g., feature values) are continuous and the (hidden) state is $\{W = w\}$, we use conditional probability density function $f_w : \mathbb{R}^d \to [0, \infty)$
- Interpretation: for $D \in \mathbb{R}^d$,

$$\mathbb{P}\left(\mathbf{X} \in D | W = w\right) = \int_D f_w(\mathbf{x})\ d\mathbf{x}$$

- Example: Traffic volume on different roads at 7:30 AM

$$X = \text{traffic on Route 1(going south)}$$
$$Y = \text{traffic on 295 (going south))}$$

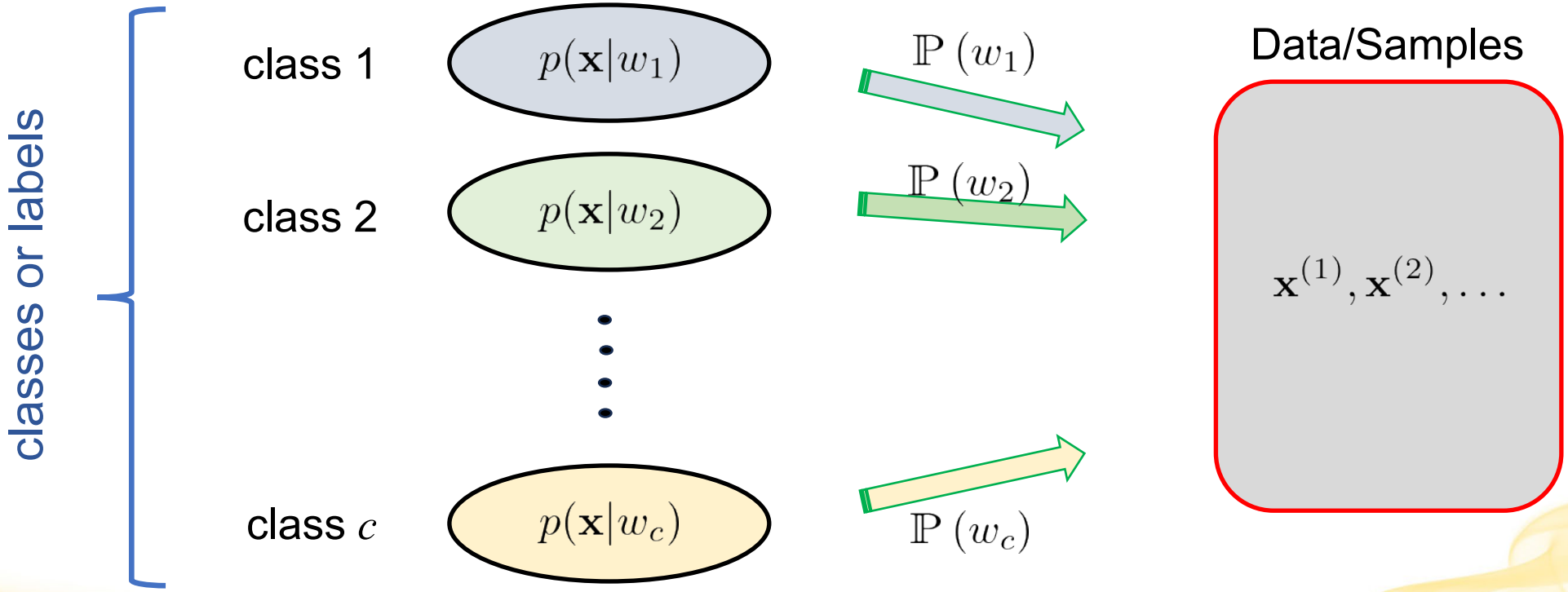
$f_N(X, Y)$ - conditional PDF on a normal weekday (N)
$f_W(X, Y)$ - conditional PDF on weekend (W)
$f_H(X, Y)$ - conditional PDF on holiday (H)

BAYESIAN DECISION THEORY

DATA/MSML 603: Principles of Machine Learning

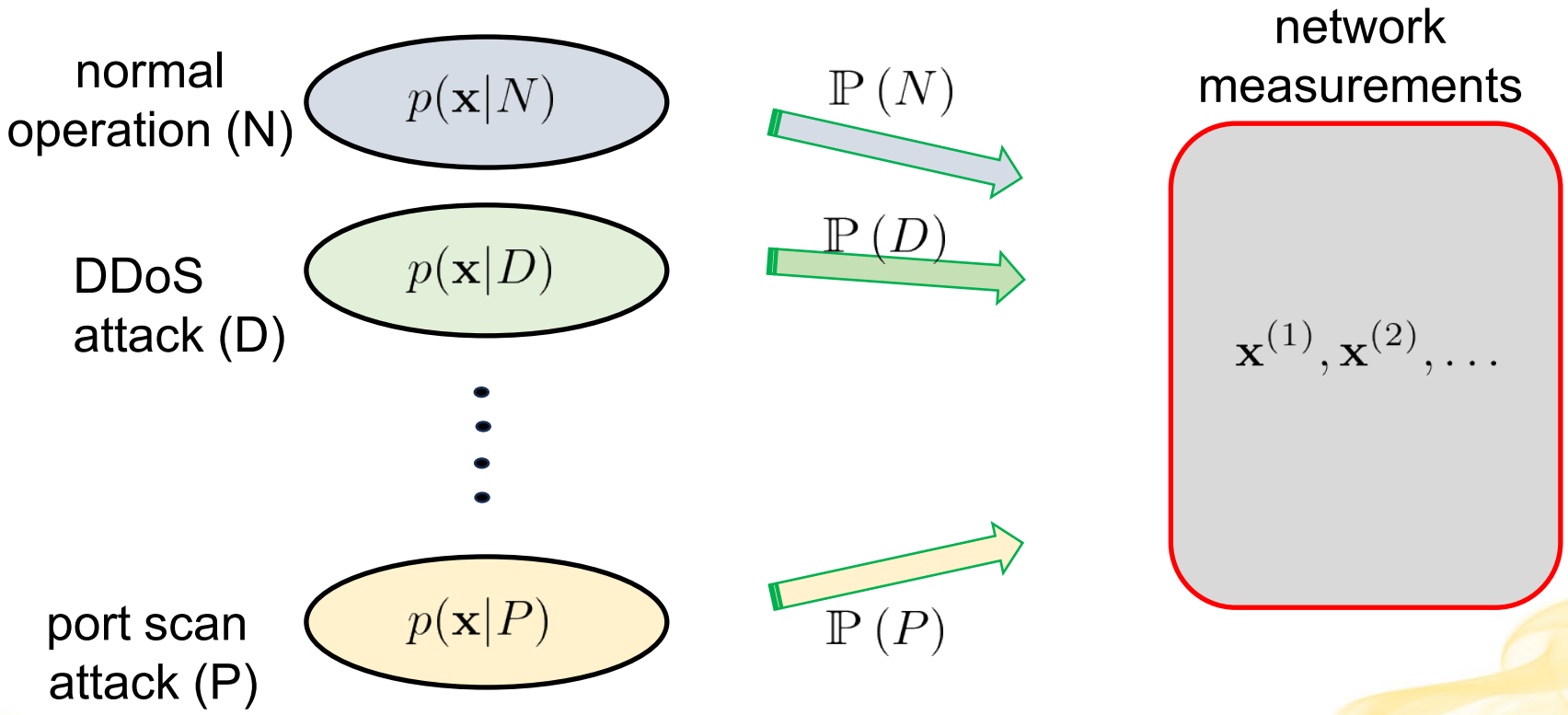
Simple Picture

- In many cases, we will be working with the following setup



Simple Picture

- Example: Network measurements during normal operation and different types of attacks



Structure of what is to come

- We will start with the simplest scenario where all necessary probabilistic information and cost/loss information is known
  - Optimal decision rules can be designed
- Then, we will relax/remove some of probabilistic information and instead estimate it from samples
  - Parameter estimation
- Finally, we will move to scenarios where little to no probabilistic information is assumed to be available

Simple Example (Classification)

- Each fish caught is one of two types – salmon (S) or tuna (T)

$$\mathbb{P}(S) = q_S, \ \mathbb{P}(T) = q_T = 1 - q_S$$

priors

  - Assumption: true independently of previously caught fish

- Scenario 1: Suppose that the priors are known, but no other information is provided. We would like to guess the type of next fish we catch before seeing it or getting any other information about it. How do we minimize the probability of guessing incorrectly?

Simple Example (Classification)

- In the absence of any additional information, we need to guess based only on the "priors"

$$\mathbb{P}(S) = q_S, \ \mathbb{P}(T) = q_T = 1 - q_S$$

- Example: Suppose $\mathbb{P}(S) = 0.4, \ \mathbb{P}(T) = 0.6$
  - Intuitively, we should choose "tuna" each time to minimize the probability of incorrect guess
- If we choose "salmon" with probability θ_S and "tuna" with probability $\theta_T = 1 - \theta_S$

$$\begin{aligned}\mathbb{P}(\text{error}) &= \mathbb{P}(\text{error} \cap S) + \mathbb{P}(\text{error} \cap T) = \mathbb{P}(\text{error} \mid S)\mathbb{P}(S) + \mathbb{P}(\text{error} \mid T)\mathbb{P}(T) \\ &= (1 - \theta_S)\mathbb{P}(S) + \theta_S\mathbb{P}(T) = \mathbb{P}(S) + \theta_S(\mathbb{P}(T) - \mathbb{P}(S))\end{aligned}$$

Simple Example (Classification)

- Scenario 2: Suppose that the priors are known. In addition, someone told us the weight of the fish but no other information is provided. How can we minimize the probability of incorrect guess?
  - Knowing the weight of the fish reveals some information about the fish. How do we use the information?
  - Remark: The information revealed to us could be a discrete attribute instead, such as its color (e.g., light orange, pink, dark red, maroon)

Simple Example (Classification)

- Scenario 2: Suppose that the priors are known. In addition, someone told us the weight of the fish but no other information is provided. How can we minimize the probability of incorrect guess?
  - In Scenario 1, we had to base our guess solely on priors (a priori probabilities)
  - Those priors told us the "likelihoods" of a fish being a salmon or a tuna before we were told the weight
  - We should "update" our estimate of likelihoods using the weight information
  - So, what should we replace priors with?
  - Answer: Posterior probabilities

Simple Example (Classification)

- Scenario 2: Suppose that the priors are known. In addition, someone told us the weight of the fish but no other information is provided. How can we minimize the probability of incorrect guess?
- Posterior probabilities: conditional probability that the fish is a salmon or a tuna after we found out its weight

$$\mathbb{P}(S \mid \text{Weight} = w) \quad \text{and} \quad \mathbb{P}(T \mid \text{Weight} = w)$$

- Question: How do we calculate these posterior probabilities?

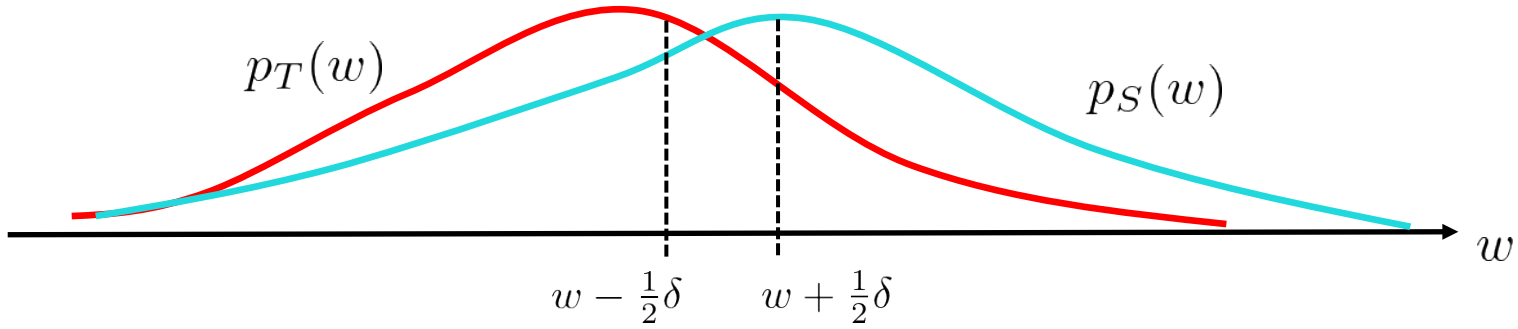
Simple Example (Classification)

- Conditional probability density functions:

$$p(w \mid S) = p_S(w) \quad \text{and} \quad p(w \mid T) = p_T(w), \qquad w \in \mathbb{R}_{++} := (0, \infty)$$

- Interpretation:

$$\mathbb{P}\left(\text{weight} \in (w - \tfrac{1}{2}\delta, w + \tfrac{1}{2}\delta) \mid S\right) \approx p_S(w) \cdot \delta \quad \text{and} \quad \mathbb{P}\left(\text{weight} \in (w - \tfrac{1}{2}\delta, w + \tfrac{1}{2}\delta) \mid T\right) \approx p_T(w) \cdot \delta$$



Simple Example (Classification)

- Approximate posterior probability:

$$\mathbb{P}\left(T \mid \text{weight} \in (w-0.5\delta, w+0.5\delta)\right) = \frac{\mathbb{P}\left(\text{weight} \in (w-0.5\delta, w+0.5\delta) \mid T\right)\mathbb{P}(T)}{\mathbb{P}\left(\text{weight} \in (w-0.5\delta, w+0.5\delta)\right)}$$

where $\mathbb{P}\left(\text{weight} \in (w-0.5\delta, w+0.5\delta)\right) = \mathbb{P}\left(\text{weight} \in (w-0.5\delta, w+0.5\delta) \mid S\right)\mathbb{P}\left(S\right) + \mathbb{P}\left(\text{weight} \in (w-0.5\delta, w+0.5\delta) \mid T\right)\mathbb{P}\left(T\right)$

- Posterior probability obtained as the limit as δ goes to 0

$$\mathbb{P}\left(T \mid \text{weight} = w\right) = \frac{p_T(w)\mathbb{P}\left(T\right)}{p(w)} \quad \text{and} \quad \mathbb{P}\left(S \mid \text{weight} = w\right) = \frac{p_S(w)\mathbb{P}\left(S\right)}{p(w)}$$

where $p(w) = p_S(w)\mathbb{P}\left(S\right) + p_T(w)\mathbb{P}\left(T\right)$

Simple Example (Classification)

- (Following the same line of reasoning) If we choose “salmon” with probability θ_S and “tuna” with probability $\theta_T = 1 - \theta_S$

$$\begin{aligned}\mathbb{P}(\text{error} \mid \text{weight} = w) &= \mathbb{P}(\text{error} \mid S, \text{weight} = w)\mathbb{P}(S \mid \text{weight} = w) \\ &\quad + \mathbb{P}(\text{error} \mid T, \text{weight} = w)\mathbb{P}(T \mid \text{weight} = w) \\ &= (1 - \theta_S)\mathbb{P}(S \mid \text{weight} = w) + \theta_S \mathbb{P}(T \mid \text{weight} = w) \\ &= \mathbb{P}(S \mid \text{weight} = w) + \theta_S \big(\mathbb{P}(T \mid \text{weight} = w) - \mathbb{P}(S \mid \text{weight} = w)\big)\end{aligned}$$

Choose Tuna if $\big(\mathbb{P}(T \mid \text{weight} = w) - \mathbb{P}(S \mid \text{weight} = w)\big) > 0$

Otherwise, choose Salmon

Simple Example (Classification)

- Comment: The denominator in the posterior probabilities, called the “evidence”, are identical

$$\mathbb{P}\left(B \mid \text{weight} = w\right) = \frac{p_T(w)\mathbb{P}\left(T\right)}{p(w)} \quad \text{and} \quad \mathbb{P}\left(S \mid \text{weight} = w\right) = \frac{p_S(w)\mathbb{P}\left(S\right)}{p(w)}$$

  - Can use the quantities in the numerators for comparison without worrying about the denominator

- If additional information is revealed, we update the posterior probabilities again to improve our “guess”

More Formal Framework (Classification)

- Framework
  1. Set of possible states $\{w_1, w_2, \ldots, w_c\}$ of nature
  2. Set of possible actions $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_a\}$
  3. Cost/loss function: $\lambda(\alpha_i \mid w_j)$ - cost/loss incurred for taking action α_i when the state of nature is w_j
  4. Feature vector $\mathbf{x} = (x_1, \ldots, x_d)$: conditional probability density functions

$$p(\mathbf{x}|w_j), j = 1, \ldots, c$$

  5. Priors: $\mathbb{P}(w_j), j = 1, \ldots, c$

More Formal Framework (Classification)

- Posterior probabilities: $$\mathbb{P}(w_j \mid \mathbf{x}) = \frac{p(\mathbf{x}|w_j)\mathbb{P}(w_j)}{p(\mathbf{x})}, \; j = 1, \ldots, c$$

  where $p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|w_j)\mathbb{P}(w_j)$

- Expected cost/loss associated with taking action α_i when $\{\mathbf{X} = \mathbf{x}\}$

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|w_j)\mathbb{P}(w_j|\mathbf{x})$$

  - Called "conditional risk"

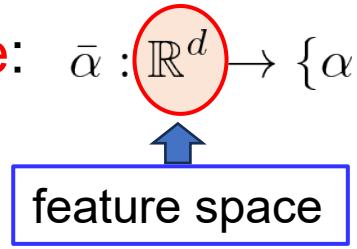
- Example: loan approval for applicants $\{w_1 = \text{good}, \; w_2 = \text{bad}\}$

$$\mathbf{X} = (\text{income, asset, education level, age, credit score})$$

$$\mathcal{A} = \{\text{approve, deny}\} \Rightarrow R(\text{approve}|\mathbf{x}), \; R(\text{deny}|\mathbf{x})$$

More Formal Framework (Classification)

- Decision rule: $\bar{\alpha} : \mathbb{R}^d \rightarrow \{\alpha_1, \ldots, \alpha_a\}$ - chooses an action based on feature vector



- Overall risk for decision rule $\bar{\alpha} : \mathbb{R}^d \rightarrow \{\alpha_1, \ldots, \alpha_a\}$

$$R^{\bar{\alpha}} = \int_{\mathbb{R}^d} R(\bar{\alpha}(\mathbf{x})|\mathbf{x})\ p(\mathbf{x})\ d\mathbf{x}$$

(or Bayes classifier)

- Bayesian decision rule: Given $\mathbf{x} \in \mathbb{R}^d$, choose action with minimum conditional risk

$$\alpha_B(\mathbf{x}) \in \arg\min_{\alpha \in \mathcal{A}} R(\alpha|\mathbf{x}) \qquad \left(\text{where } R(\alpha|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha|w_j)\mathbb{P}\left(w_j|\mathbf{x}\right)\right)$$

  - Bayes risk R^{α_B} is the smallest overall risk we can achieve using any decision rule

➡ Optimal decision rule/policy

Discriminant Functions for Classification

- Representation of classifier/decision rule

1. Discriminant functions $g_i, i \in \{1, \ldots, c\} =: \mathcal{C}$ ⬅ label set
   - Assign a feature vector $\mathbf{x} \in \mathbb{R}^d$ to class w_i if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \quad \text{for all } j \neq i$$

   - Example 1: Bayes classifier - $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$ (minus conditional risk)
   - Example 2: Minimum error-rate classifier

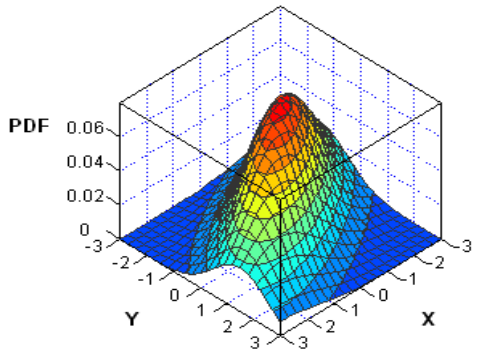
$$g_i(\mathbf{x}) = \mathbb{P}(w_i|\mathbf{x}) = \frac{p(\mathbf{x}|w_i)\mathbb{P}(w_i)}{\sum_{j=1}^{c} p(\mathbf{x}|w_j)\mathbb{P}(w_j)}$$

$$\Leftrightarrow \quad g_i(\mathbf{x}) = \log\left(p(\mathbf{x}|w_i)\right) + \log\left(\mathbb{P}(w_i)\right) \qquad \text{(will be used in Gaussian case)}$$

Discriminant Functions for Classification

- Decision rule partitions the feature space into c decision regions $\mathcal{R}_1, \ldots, \mathcal{R}_c$
  - Separated by decision boundaries surface in feature space where ties occur among the largest discriminant functions

Gaussian Case



- Special case: Suppose conditional probability density function (PDF) of feature vector is given by

$$p(\mathbf{x}|w_i) = \frac{1}{(2\pi)^{d/2}\sqrt{|\Sigma_i|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right), \quad i = 1, 2, \ldots, c$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^d$, $\Sigma_i \in \mathbb{R}^{d\times d}$ (positive definite in most cases)

- Discriminant functions

Independent of states w_i ➡ **can be dropped**

$$\begin{aligned} g_i(\mathbf{x}) &= \log\left(p(\mathbf{x}|w_i)\right) + \log\left(\mathbb{P}\left(w_i\right)\right) \\ &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log\left(|\Sigma_i|\right) + \log(\mathbb{P}\left(w_i\right)) \\ &= \log\left(p(\mathbf{x}|w_i)\right) \end{aligned}$$

Gaussian Case: Case 1

- Independent feature values: $\Sigma_i = \sigma^2 \mathbf{I}$ $(\sigma^2 > 0)$ ⇒ $|\Sigma_i| = \sigma^{2d}$ (product of eigenvalues), $\Sigma_i^{-1} = \sigma^{-2}\mathbf{I}$

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\|\mathbf{x} - \boldsymbol{\mu}_i\|^2 + \log(\mathbb{P}(w_i)) = -\frac{1}{2\sigma^2}\left(\|\mathbf{x}\|^2 - 2\boldsymbol{\mu}_i^T\mathbf{x} + \|\boldsymbol{\mu}_i\|^2\right) + \log(\mathbb{P}(w_i))$$

(ignoring common terms)

independent of states

- Linear discriminant functions

$$g_i(\mathbf{x}) = \underbrace{\frac{1}{\sigma^2}\boldsymbol{\mu}_i^T}_{=\mathbf{w}_i^T}\mathbf{x} \underbrace{- \frac{1}{2\sigma^2}\|\boldsymbol{\mu}_i\|^2 + \log(\mathbb{P}(w_i))}_{=w_{i0}} = \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

"threshold" or "bias"

⇒ Decision boundaries given by hyperplanes

Gaussian Case: Case 2

- Identical covariance matrix: $\Sigma_i = \Sigma \quad \text{for all } i$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log\left(|\Sigma|\right) + \log(\mathbb{P}(w_i))$$

independent of states

- Remove common terms and quadratic term

$$g_i(\mathbf{x}) = \underbrace{\left(\Sigma^{-1}\boldsymbol{\mu}_i\right)^T}_{=\mathbf{w}_i^T}\mathbf{x} \underbrace{- \frac{1}{2}\boldsymbol{\mu}_i^T\Sigma^{-1}\boldsymbol{\mu}_i + \log(\mathbb{P}(w_i))}_{=w_{i0}} = \mathbf{w}_i^T\mathbf{x} + w_{i0}$$

linear discriminant functions

→ Decision boundaries given by hyperplanes

Gaussian Case: Case 2

- Boundary between $\mathcal{R}_i$ and $\mathcal{R}_j$ can be obtained by solving a system of linear equations: $g_i(\mathbf{x}) = g_j(\mathbf{x})$

Gaussian Case: Case 3

- Arbitrary positive definite covariance matrices Σ_i

$$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + w_{i0}$$

where $\mathbf{W}_i = -\frac{1}{2}\Sigma_i^{-1},\ \mathbf{w}_i = \Sigma_i^{-1}\boldsymbol{\mu}_i,\ w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^T\Sigma_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\log(|\Sigma_i|) + \log(\mathbb{P}(w_i))$

  - Can lead to complicated decision regions and boundaries