EECE 5644: Bayesian Decision Theory

Mark Zolotas

E-mail: m.zolotas@northeastern.edu
Webpage: https://coe.northeastern.edu/people/zolotas-mark/

Tentative Course Outline (Wks. 1-2)

Topics	Dates	Assignments	Additional Reading
~~Course Overview~~ ~~Machine Learning Basics~~	~~07/05~~	**Optional Homework 0** released on Canvas on 07/08 but please do NOT submit on Canvas	~~Chpt. 1~~ ~~Murphy 2012~~
~~Foundations: Linear Algebra, Probability, Numerical Optimization (Gradient Descent), Regression~~	~~07/06-12~~		~~Stanford LA Review~~ ~~Stanford Prob. Review~~ ~~Chpt. 8 Murphy 2022~~
~~*Quick Python Tutorial*~~	~~07/12~~	**Homework 1** released on Canvas on 07/15 **Due 07/25**	~~N/A~~
~~Linear Classifier Design, Linear Discriminant Analysis~~ and Principal Component Analysis (PCA)	07/13-15		~~Chpts. 9.2~~ & 20.1 Murphy 2022
Bayesian Decision Theory: Empirical Risk Min, Max Likelihood (ML), Max a Posteriori	07/14		Chpt. 2 Duda & Hart 2001 Deniz Erdogmus Notes

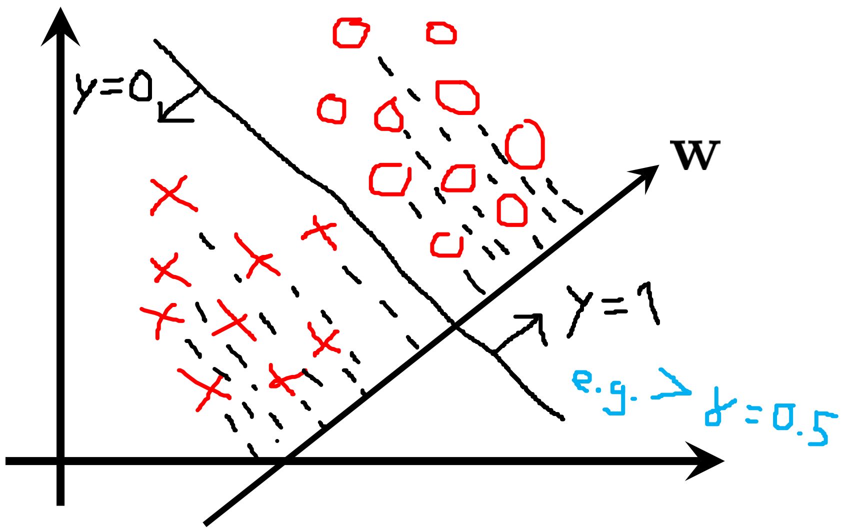
Recap: Fisher's LDA Classifier

Let $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$, N training samples
Inputs $\mathbf{x} \in \mathbb{R}^n$, binary valued labels $y \in \{0, 1\}$

- Given Fisher's solution: $\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{w}_{\text{LDA}} = \lambda\mathbf{w}_{\text{LDA}}$
- **Decision rule** based on Fisher's LDA projection:

$$\mathbf{w}_{\text{LDA}}^{\mathsf{T}}\mathbf{x} \underset{\hat{y}=0}{\overset{\hat{y}=1}{\gtrless}} \gamma$$

- Can decide γ threshold using **ROC curves**

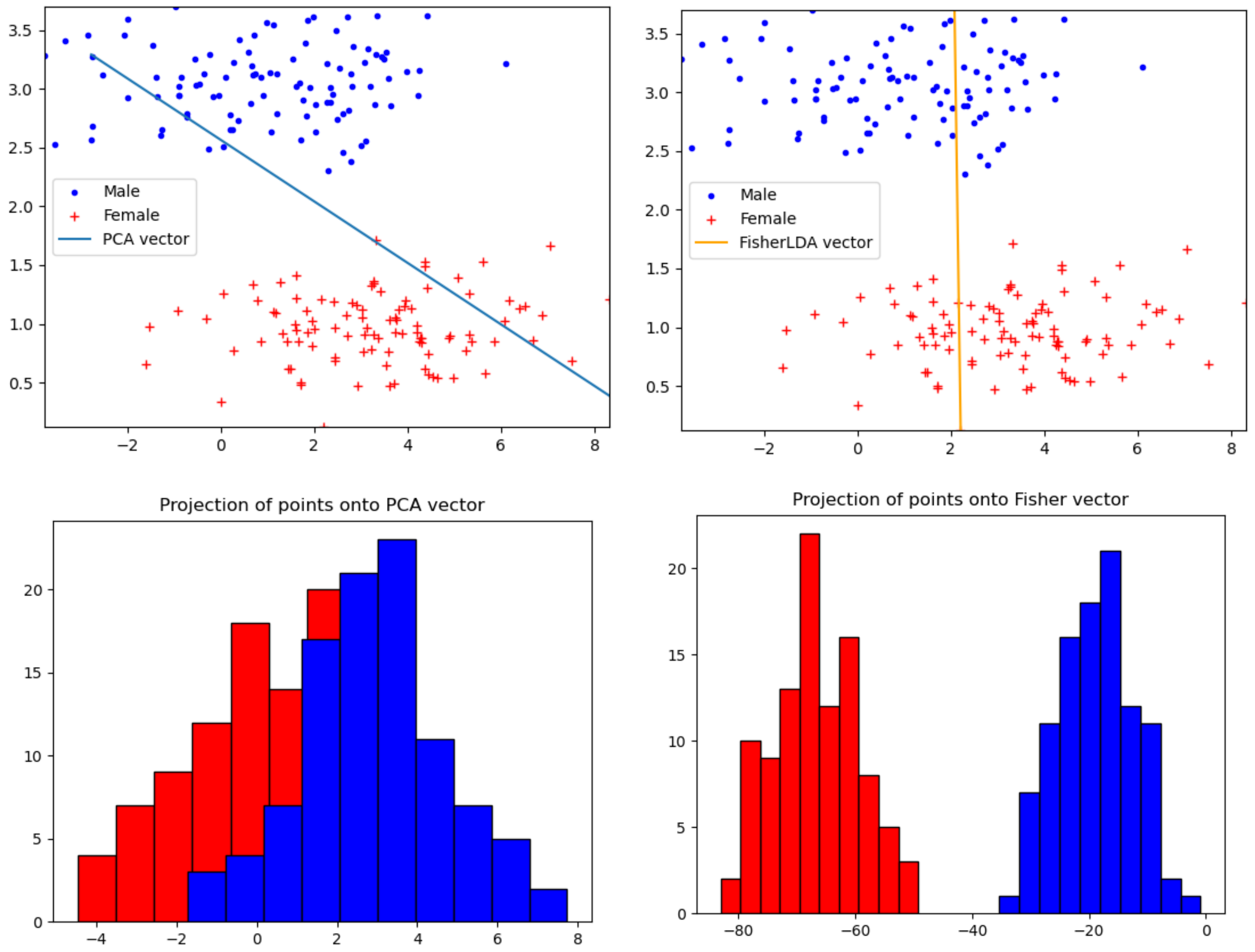


Hyperplane
in n-dims.

Recap: Data Representation vs Classification

- **Data Representation:** Project data to lower dimensional space that *most accurately represents* the original data, e.g., PCA projects in directions of maximum variance

- **Data Classification:** Project data to a low dimensional space that preserves structure useful for classification, e.g., Fisher's LDA



Kevin Murphy, *"Probabilistic Machine Learning: An Introduction"*, 2022

Probabilistic Machine Learning

- Classification in a functional input-output way:

$$y = f(\mathbf{x}; \boldsymbol{\theta})$$

- Cannot perfectly predict input-output mappings, there is always **uncertainty**
  - **Epistemic/Model:** From limited knowledge of f, e.g., not enough data
  - **Aleatoric/Data:** From intrinsic randomness in the data, e.g., noisy input source

- Must capture uncertainty in the classification; think **conditional probability**:

$$p(L = y \,|\, \mathbf{x}, \boldsymbol{\theta}) = f_y(\mathbf{x}; \boldsymbol{\theta})$$

- Now $f_y(\mathbf{x}; \boldsymbol{\theta})$ returns the probability of class label y from all labels L

Statistical Approach

- **Statistical methods** in machine learning assume that whatever process "generating" our data is governed by rules of probability

Let $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$ be our dataset of N random vector samples $\mathbf{x}^{(i)} \in \mathbb{R}^n$

- Also assume samples are **independent & identically distributed (iid)**:

Product of marginals

$$p(\mathcal{D}) = p_{X_1,\ldots,X_N}(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}) = p_{X_1}(\mathbf{x}^{(1)}) \cdots p_{X_N}(\mathbf{x}^{(N)})$$

Joint likelihood of entire dataset

$$= p_X(\mathbf{x}^{(1)}) \cdots p_X(\mathbf{x}^{(N)})$$

Identical RV

- Under this probabilistic framework, how can we make classification decisions using (**posterior**) probability $p(L = y \mid \mathbf{x})$?

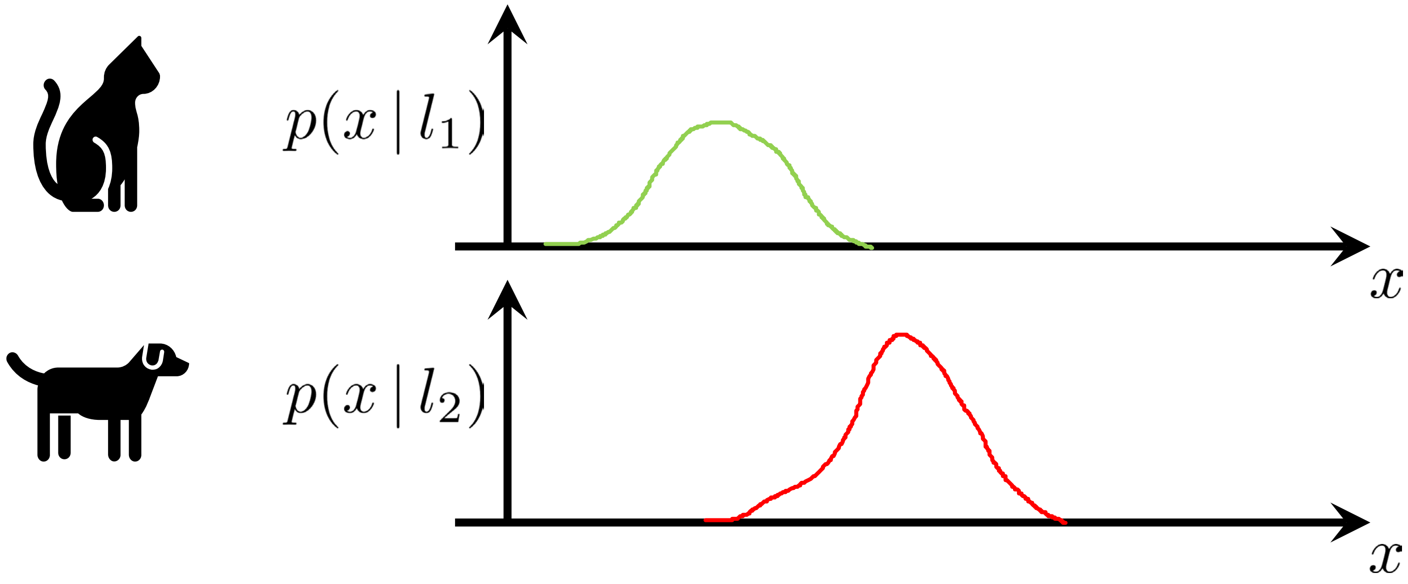
Example: Image Classification of Dogs & Cats

- **Goal:** Decide whether an unseen image is either a cat or dog
- In decision-theoretic terminology:
  - ❖ Our **hypotheses** are our labels L, where $L = 1$ for cat and $L = 2$ for dog
  - ❖ Our **decisions** D in this setting also correspond to our labels L

Class Conditional Probabilities

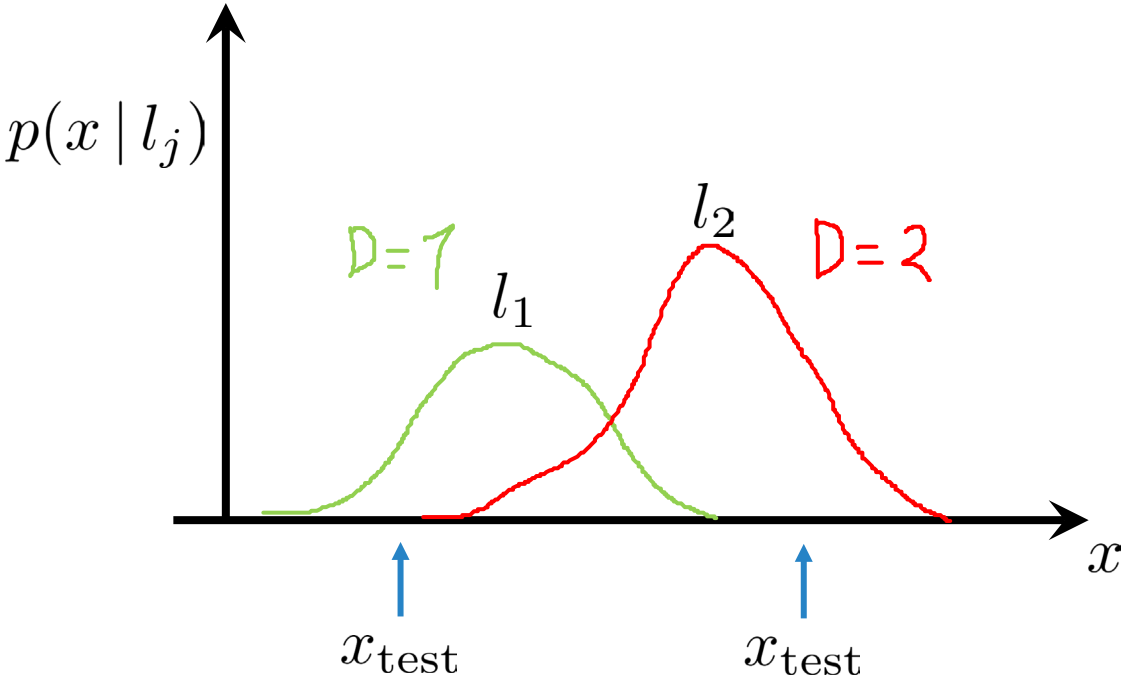
- Probabilities $p(\mathbf{x} \mid L = j)$ of observations $\mathbf{x}$ whose distribution depends on a particular hypothesis or label j, also denoted as $p(\mathbf{x} \mid l_j)$
- This distribution is known as the **class-conditional likelihood**
- E.g. $\mathbf{x}$ is a random vector of variables, like color features, tail-to-body ratio...

Look at one scalar variable $x \in \mathbb{R}$



Decide Off Likelihood

- Need a decision rule for whether x_{test} is from the cat or dog class



- Deciding labels that assign higher **likelihood** $p(x \mid l_j)$

Maximum Likelihood (ML) Classification

- For two classes and $\mathbf{x} \in \mathbb{R}^n$, the decision rule $D(\mathbf{x})$ could be a **ratio of likelihoods**:

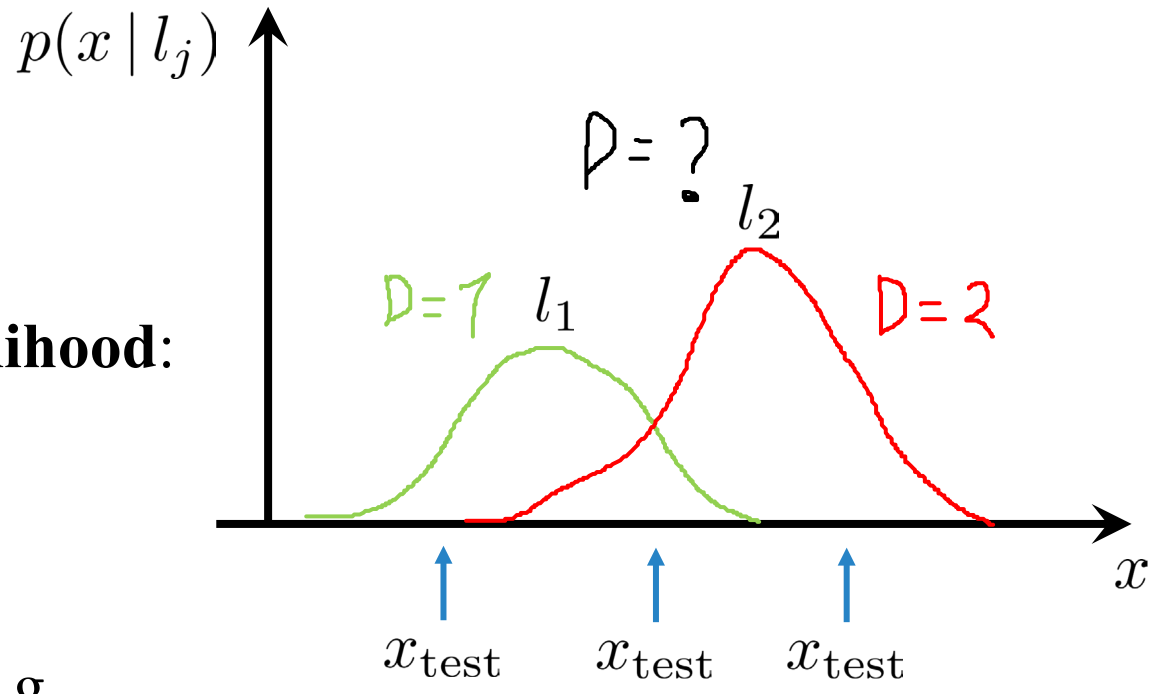
$$\frac{p(\mathbf{x}\,|\,l_2)}{p(\mathbf{x}\,|\,l_1)} \underset{D(\mathbf{x})=1}{\overset{D(\mathbf{x})=2}{\gtrless}} 1$$

- For $C > 2$, choose **maximum likelihood**:

Decision Rule

$$\boxed{D(\mathbf{x}) = \underset{j\in\{1,\ldots,C\}}{\arg\max}\; p(\mathbf{x}\,|\,l_j)}$$

- Or what if one class if very rare? E.g., only a few images of dogs

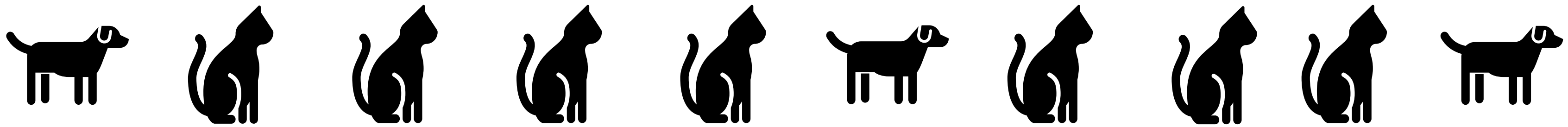


Class Priors

- If we had an equal number of cat and dog images, we would think the next new image encountered is equally likely to be a cat or dog

- **Class Prior:** Assumed *a priori* probability $p(L = j)$ of a data point belonging to a particular class j

- E.g., our dataset reflects prior knowledge of how likely we are to see cat/dog

$\mathcal{D} =$ 

$$p(l_1) = 0.7 \qquad p(l_2) = 0.3 \qquad \sum_j p(l_j) = 1$$

Decide Off Prior Knowledge

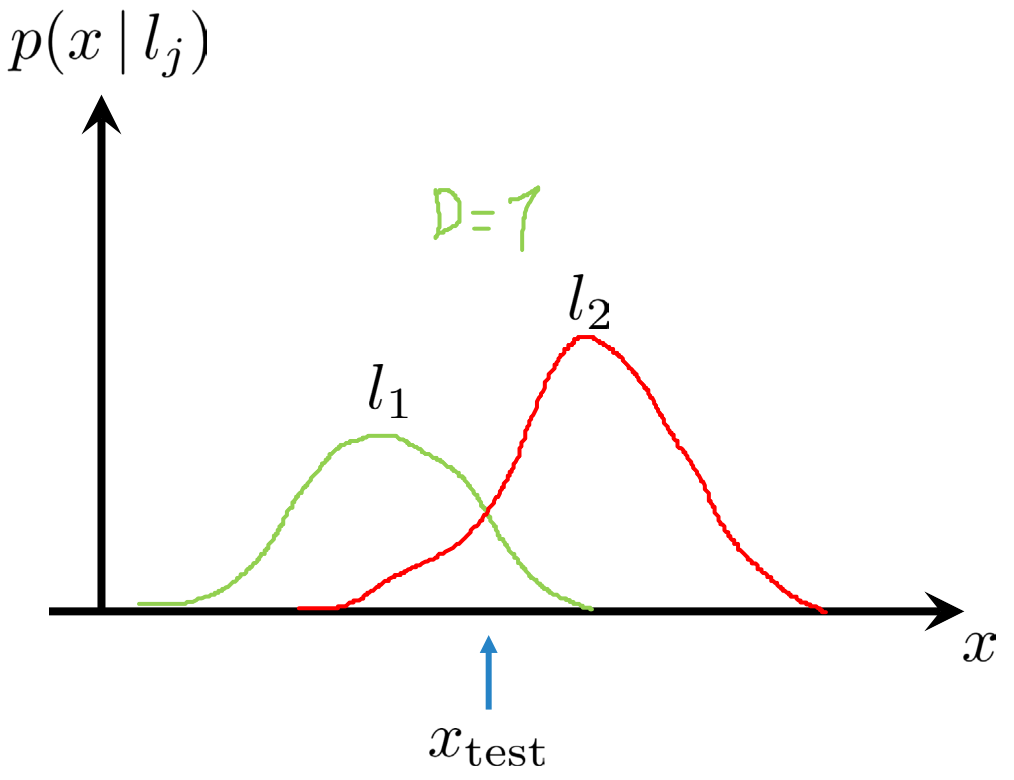
- Assume incorrect classifications of cats/dogs have the **same cost** or effect
- If the only information we could use is based on our prior probabilities, then in this uninformed state we could use the following **decision rule**:

Decide cat (1) if $p(l_1) > p(l_2)$; else dog (2)

- Probably not a good idea as we would **repeatedly** make the **same decision** even though both types of images might appear in unseen data
- Better to make use of the **information/evidence** we have available

Decide Off Prior AND Likelihood

How do we decide a class now?



$p(l_1) = 0.7$

$p(l_2) = 0.3$

- Given these prior probabilities AND likelihood, would decide cat (1)

Bayes' Theorem

- Interested in the joint probability:

$$p(\mathbf{x}, l_j) = p(\mathbf{x} \,|\, l_j)p(l_j) = p(l_j \,|\, \mathbf{x})p(\mathbf{x})$$

- **Bayes' Theorem** lets us relate these conditional distributions:

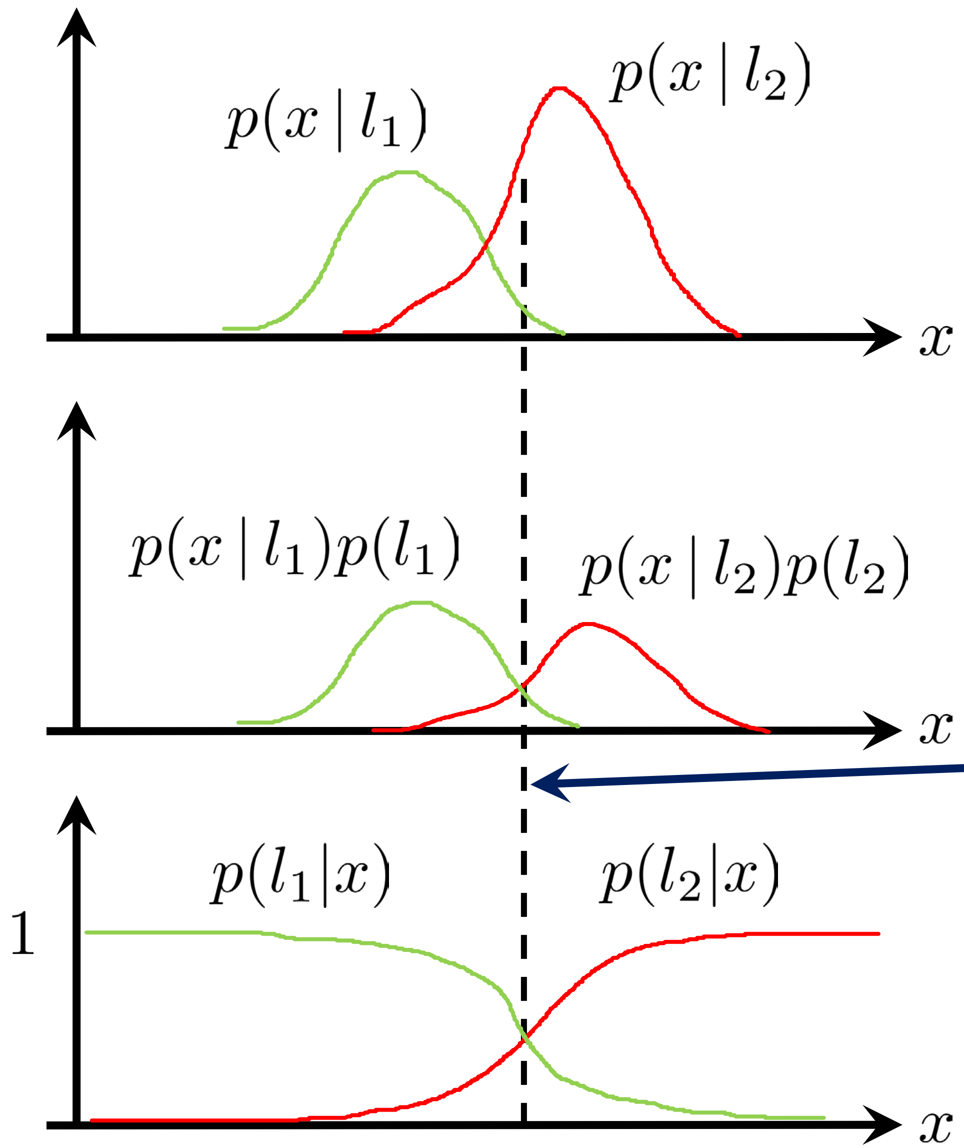
$$\boxed{p(l_j \,|\, \mathbf{x}) = \frac{p(\mathbf{x} \,|\, l_j)p(l_j)}{p(\mathbf{x})}}$$

Normalization Constant

- In plain English:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Posterior Probability



likelihood

likelihood $\times$ prior

Decision Boundary

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Bayes' Decision Theory – Min. Probability of Error Rule

- **Goal:** Minimize **misclassification rate** or **probability of error**
- For $\underline{C = 2}$: $$\Pr(\text{error}) = \begin{cases} p(l_1 \mid \mathbf{x}) & \text{if decide } l_2 \\ p(l_2 \mid \mathbf{x}) & \text{if decide } l_1 \end{cases}$$
- So decide cat (1) if $p(l_1 \mid \mathbf{x}) > p(l_2 \mid \mathbf{x})$; else dog, i.e., $\Pr(\text{error}) = \min_{j \in \{1,2\}} p(l_j \mid \mathbf{x})$
- Equivalently, decide 1 if: $$\frac{p(\mathbf{x} \mid l_1) p(l_1)}{\cancel{p(\mathbf{x})}} > \frac{p(\mathbf{x} \mid l_2) p(l_2)}{\cancel{p(\mathbf{x})}}$$

Const. wrt l

$$p(\mathbf{x} \mid l_1) p(l_1) > p(\mathbf{x} \mid l_2) p(l_2)$$

Optimal Bayes Classifier $\Longrightarrow$ $$\boxed{\frac{p(\mathbf{x} \mid l_1)}{p(\mathbf{x} \mid l_2)} > \frac{p(l_2)}{p(l_1)}}$$

Bayes' Decision Theory – Max. a Posteriori (MAP)

- **"Optimal"** but assumes relevant probability terms, e.g., $p(\mathbf{x} \mid l_j)$ are known

$$\frac{p(\mathbf{x} \mid l_1)}{p(\mathbf{x} \mid l_2)} > \frac{p(l_2)}{p(l_1)}$$

*Rarely know true probabilities; **fit models** instead by estimating $\boldsymbol{\theta}$*

- Special Cases:
  - If $\mathbf{x}$ uninformative about labels, $p(\mathbf{x} \mid l_1) = p(\mathbf{x} \mid l_2)$ then decide on <u>priors</u>
  - If $\mathbf{x}$ uniformly distributed, i.e., $p(l_1) = p(l_2)$ then decide on <u>likelihoods</u>

If uniform then MAP equivalent to ML

- **Decision Rule** equivalence to **maximum a posteriori**, so for $\underline{C > 2}$ classes:

$$\boxed{D(\mathbf{x}) = \underset{j \in \{1,\ldots,C\}}{\arg\max}\ p(l_j \mid \mathbf{x})} = \underset{j \in \{1,\ldots,C\}}{\arg\max}\ \underline{p(\mathbf{x} \mid l_j)}p(l_j)$$

Empirical Risk Minimization (ERM) – Motivation

- Generalize:
  - ❖ Allow **actions** D that are not just deciding classes/labels L, e.g., "rejection"
  - ❖ Introduce a **loss/cost function** more general than the probability of error, e.g., cases where classification errors are <u>not all equal</u>

- Examples:
  - ❖ Must be certain that patient is sick before reporting diagnosis
  - ❖ Treatment plans have side-effects and trade-off costs depending on the patient
  - ❖ Reporting a fire is vital and so false alarms are acceptable (less *risky*)

Empirical Risk Minimization (ERM) – Terminology

- Notation:
  - Let $L = \{1, \dots, C\}$ be the finite set of **labels** ("states of nature")
  - Let $D = \{1, \dots, A\}$ be the finite set of possible **actions/decisions**
  - Then $\boldsymbol{\Lambda}$ is the **loss matrix** such that λ_{ij} is the **loss/cost** associated with deciding action i when the true label is j

$$\Lambda = \begin{array}{cc} i & \begin{bmatrix} & \vert & \\ - & \lambda_{ij} & - \\ & \vert & \end{bmatrix} \\ & j \end{array} \in \mathbb{R}^{A \times C}$$

- Introduce notion of **risk** as **expected loss/cost** for a decision rule $D(\boldsymbol{x})$:

$$\mathbb{E}_X[R] = \int_{-\infty}^{\infty} R(D(\mathbf{x}) \,|\, \mathbf{x}) p_X(\mathbf{x}) d\mathbf{x}$$

Empirical Risk Minimization (ERM) – Formulation

$$\mathbb{E}_X[R] = \int_{-\infty}^{\infty} R(D(\mathbf{x}) \,|\, \mathbf{x}) p_X(\mathbf{x}) d\mathbf{x}$$

- **Conditional risk** of taking an action/decision $D(\mathbf{x}) = i$ for a given $\mathbf{x}$:

d_i

$$R(\underline{D(\mathbf{x}) = i} \,|\, \mathbf{x}) = \sum_{j=1}^{C} \lambda_{ij} p(\underline{L = j} \,|\, \mathbf{x})$$

L_j

- **Empirical Risk Minimization:** $\min_i R(D(\mathbf{x}) = i \,|\, \mathbf{x})$

- **ERM Decision Rule:** $\boxed{D(\mathbf{x}) = \arg\min_i R(D(\mathbf{x}) = i \,|\, \mathbf{x})}$

Two Category Classification – Setting

- Two categories l_j and two decisions d_i for $i, j \in \{1, 2\}$
- Conditional risk for each decision:
$$R(d_1 \,|\, \mathbf{x}) = \lambda_{11} p(l_1 \,|\, \mathbf{x}) + \lambda_{12} p(l_2 \,|\, \mathbf{x})$$
$$R(d_2 \,|\, \mathbf{x}) = \lambda_{21} p(l_1 \,|\, \mathbf{x}) + \lambda_{22} p(l_2 \,|\, \mathbf{x})$$
- Per ERM, decide $D = 1$ if $R(d_1 \mid x) < R(d_2 \mid x)$ and vice versa
- The ERM decision rule in this case is thus:

$$\boxed{\lambda_{11} p(l_1 \,|\, \mathbf{x}) + \lambda_{12} p(l_2 \,|\, \mathbf{x}) \underset{D(\mathbf{x}) = 1}{\overset{D(\mathbf{x}) = 2}{\gtrless}} \lambda_{21} p(l_1 \,|\, \mathbf{x}) + \lambda_{22} p(l_2 \,|\, \mathbf{x})}$$

Two Category Classification – Intuition

$$\lambda_{11}p(l_1 \,|\, \mathbf{x}) + \lambda_{12}p(l_2 \,|\, \mathbf{x}) \underset{D(\mathbf{x})=1}{\overset{D(\mathbf{x})=2}{\gtrless}} \lambda_{21}p(l_1 \,|\, \mathbf{x}) + \lambda_{22}p(l_2 \,|\, \mathbf{x})$$

- Rearrange:

$$(\lambda_{12} - \lambda_{22})p(l_2 \,|\, \mathbf{x}) \underset{D(\mathbf{x})=1}{\overset{D(\mathbf{x})=2}{\gtrless}} (\lambda_{21} - \lambda_{11})p(l_1 \,|\, \mathbf{x})$$

- The loss incurred for **error** is usually > than the loss of being **correct**, meaning $\lambda_{12} - \lambda_{22} > 0$ and $\lambda_{21} - \lambda_{11} > 0$
- Decisions determined by the posterior probabilities scaled by loss differences

Two Category Classification – Likelihood Ratio Test

$$(\lambda_{12} - \lambda_{22})p(l_2 \,|\, \mathbf{x}) \underset{D(\mathbf{x})=1}{\overset{D(\mathbf{x})=2}{\gtrless}} (\lambda_{21} - \lambda_{11})p(l_1 \,|\, \mathbf{x})$$

- Replace posteriors by the priors and conditional densities (Bayes):

$$(\lambda_{12} - \lambda_{22})p(\mathbf{x} \,|\, l_2)p(l_2) \underset{D(\mathbf{x})=1}{\overset{D(\mathbf{x})=2}{\gtrless}} (\lambda_{21} - \lambda_{11})p(\mathbf{x} \,|\, l_1)p(l_1)$$

- Assuming $\lambda_{12} - \lambda_{22} > 0$ we can write:

$$\boxed{\frac{p(\mathbf{x} \,|\, l_2)}{p(\mathbf{x} \,|\, l_1)}} \underset{D(\mathbf{x})=1}{\overset{D(\mathbf{x})=2}{\gtrless}} \frac{(\lambda_{21} - \lambda_{11})}{(\lambda_{12} - \lambda_{22})} \frac{p(l_1)}{p(l_2)}$$

Likelihood Ratio Test: *Decide* 1 *or* 2 *based on a threshold independent of observations* $\mathbf{x}$

Zero-One Loss

- Classification problems usually view actions as decisions about labels
- For true label $L = j$, the decision $D = i$ is correct if $i = j$ and wrong if $i \neq j$
- Naturally wish to avoid errors so aim to **minimize** the **probability of error**

Error rate

- Loss function is hence **zero-one** or **symmetric** loss, where **no loss** is assigned to a correct decision, and **unit loss** to *any* error (equally costly):

As **Kronecker delta** expression $1 - \delta_{ij} \rightarrow$

$$\lambda_{ij} = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases} \qquad j, i \in \{1, ..., C\}$$

Minimum-Error-Rate Classification

- **Risk** is equivocally average error rate:

Conditional probability that d_i is correct

$$R(d_i \,|\, \mathbf{x}) = \sum_{j=1}^{C} \lambda_{ij} p(l_j \,|\, \mathbf{x}) = \sum_{j \neq i} \underbrace{p(l_j \,|\, \mathbf{x})}_{P_r(\text{error})} = 1 - \overbrace{p(l_i \,|\, \mathbf{x})}$$

Sums all j entries except i (zeroed)

- Thus to **minimize risk** is to **minimize probability of error**

- Which is identical to **maximizing posterior**: **ERM = MAP**

Decision Rule for 0-1 loss

$$\boxed{D(\mathbf{x}) = \underset{i \in \{1,\ldots,A\}}{\arg\min}\ 1 - p(l_i \,|\, \mathbf{x}) = \underset{i \in \{1,\ldots,A\}}{\arg\max}\ p(l_i \,|\, \mathbf{x})}$$

Coding Break

Concluding Remarks

- Looked at Bayesian Decision Theory and how to apply Bayes Theorem to obtain an optimal classifier
- Established decision rules based on ML, MAP, and ERM
- Code:

https://github.com/mazrk7/EECE5644_IntroMLPR_LectureCode/blob/main/notebooks/erm_decision_theory/erm_gmm.ipynb

https://github.com/mazrk7/EECE5644_IntroMLPR_LectureCode/blob/main/notebooks/erm_decision_theory/erm_decision_boundaries.ipynb

- Naïve Bayes to follow!