# ML Supervised Learning 2 by ambedkar@IISc

- Machine Learning Workflow
- Different Types of Learning
- Classification using Bayes rule
- Applications of Machine Learning

# Agenda

# Machine Learning Workflow

# Machine Learning Workflow

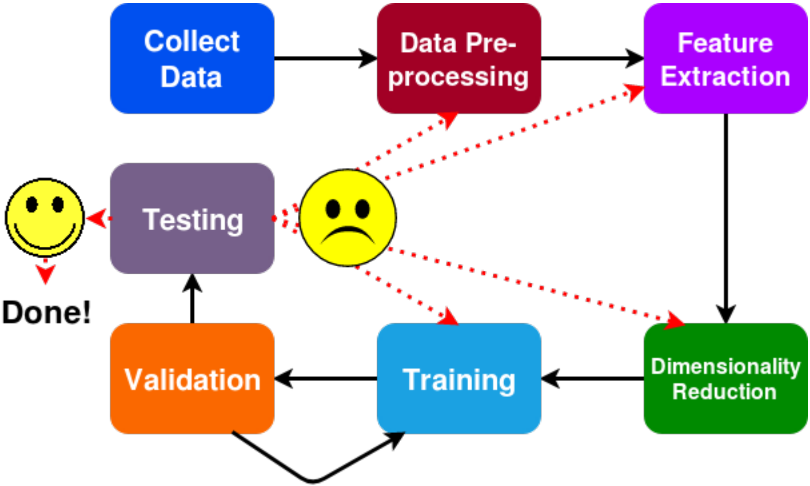
**Message is to practitioners**

There are no Royal roads to build AI systems.

The following can provide some guidance

- Mathematical foundations of machine learning.
- Systematic and scientific experimentation
- Domain knowledge
- Perseverance

# Machine Learning Workflow




*Machine Learning Workflow*

# Machine Learning Workflow: Data Pre-processing

- **Data Cleaning**
  - Removing outliers
  - Filling in missing values
  - Denoising the data
- **Normalization**
  - Making data zero mean
  - Scaling the values
- **Integration**
  - Combine data from different sources

# Machine Learning Workflow: Feature Extraction

- **Manually Finding Features**
  - Using domain expertise
  - Finding relevant information
- **Automatically Discovering Features**
  - Features themselves are learnable
  - These feature are usually not interpretable

# Machine Learning Workflow: Dimensionality Reduction

- Finding a compressed representation of data that contains approximately the same information
- Discard features that are not relevant or highly correlated
- Reduces the number of parameters needed in the model
- Leads to better generalization performance
- Use methods like Principle Component Analysis (PCA)

# Machine Learning Workflow: Other Components

- **Training**
  - Choose a model
  - Use observed data to learn parameters of the model
  - *e.g.*, learning weights of a neural network
- **Validation**
  - Use validation strategies to fine tune model hyperparameters
  - Perform model selection
  - *e.g.*, using $K$-fold cross validation to select a value of regularization parameter
- **Testing**
  - Compute the performance on unseen data
  - Diagnose the problems
  - Deploy the model

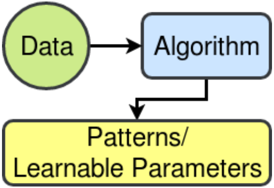
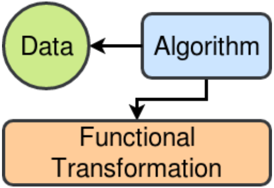
# On Learning and Different Types

# What is Learning?

It is hard to precisely define the learning problem in its full generality, thus let us consider an example:

| | **Problem 1** | **Problem 2** |
|---|---|---|
| **Input** | Some cat images $\mathbf{C} = \{C_1, C_2, \ldots, C_m\}$ and dog images $\mathbf{D} = \{D_1, D_2, \ldots, D_n\}$ | An array of numbers $\mathbf{a} = [a_1, a_2, \ldots, a_n]$ |
| **Objective** | Identify a new image $X$ as cat/dog | Sort $\mathbf{a}$ in ascending order |
| **Approach** | ? | Follow a fixed recipe that works in the same way for all arrays $\mathbf{a}$ |

# What is Learning? (contd. . . )

| Cat vs Dog | Sorting |
|---|---|
| Any approach with hard-coded "rules" is bound to fail | Hard-coded "rules" can sort any array |
| Algorithm must rely on previously observed data | Arrays sorted earlier will not affect the sorting of a new array |
| A good algorithm will get better as more data is observed | No such notion |
|   |   |

# Different types Learning Problems

- Learn by exploring data
  - Supervised Learning
  - Unsupervised Learning
- Learn from data, in a more challenging circumstances
  - Semi-supervised Learning
  - Domain Adaptation
  - Active Learning
- Learn by interacting with an environment
  - Multi-armed Bandits
  - Reinforcement Learning
- Very recent challenging AI paradigms
  - Zero/One/Few-shot Learning
  - Transfer Learning
  - Multi-agent reinforcement learning

## Classification of Learning Approaches

- Supervised Learning - Separating spam from normal emails
- Unsupervised Learning - Identifying groups in a social network

- Reinforcement Learning - Controlled medicine trials

- Zero/One/Few-shot Learning - Learning from few examples
- Transfer Learning - Multi-task learning
- Semi-supervised Learning - Using labeled and unlabeled data
- etc.

# Classification of Learning Approaches

- Supervised Learning - Separating spam from normal emails
- Unsupervised Learning - Identifying groups in a social network
- Reinforcement Learning - Controlled medicine trials
- Zero/One/Few-shot Learning - Learning from few examples
- Transfer Learning - Multi-task learning
- Semi-supervised Learning - Using labeled and unlabeled data
- etc.

# Bayesian Decision Theory

# Bayesian Decision Making in Real Life

Let us help a fisherman trying to classifying his catch. For simplicity, let us consider that he has to classify between Sea bass $(y_1)$ and Salmon $(y_2)$.

- It is a two class classification problem
- We will study this in various scenarios

# Decision Rule: Based on Prior Knowledge

Fishermen will have some domain or prior knowledge. Suppose, except for this we do not have any other knowledge.

- Suppose, in a particular season there is a more probability of catching sea bass or in a particular area probability of getting Salmon is more.
- Suppose the prior probabilities are $P(y_1)$ and $P(y_2)$. $(P(y_1) + P(y_2) = 1 \;\&\; P(y_1), P(y_2) \geq 0)$
- Rule (or common sense) says

$$\text{Decide } y_1 \text{ if } P(y_1) > P(y_2)$$
$$y_2 \text{ otherwise}$$

## Decision Rule: Based on Prior Knowledge (contd...)

How good is this?

- It looks fine but for every catch the class label is going to be the same.
- Can we feed the image of of the fish to our model so that it can consider its features before deciding on the label?

# Decision Rule: Based on class conditional probabilities

Aim here is to get features of the fish and feed it to our model.

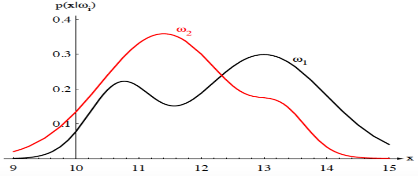
- Suppose we can get features of the fish like measurement of weight ($x$).
- We will consider the class conditional densities $P(x|y_i), \ i = \ 1, 2 \ )$, which are also called likelihood.
- $P(x|y_i)$ denotes probability of observing a particular feature(s) $x$ provided it has a class label $y_i$.

# Decision Rule: Based on class conditional probabilities (Contd...)

Now the decision Rule:

Decide $y_1$ if $P(x|y_1) > P(x|y_2)$

$y_2$ otherwise




*Likelihoods*

# Bayesian way....

Bayesian formulation helps in combining prior knowledge and class conditional probabilities into a single rule by finding posterior distribution $P(y_i|x)$

**Decision Rule: Using posterior distribution**

Using Bayes rule

$$P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)} \; i = 1, 2$$

where

$$P(x) = \sum_{i=1,2} P(x|y_i)P(y_i)$$

$P(x)$ is called evidence

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

Rule says

$y_1$ if $P(y_1|x) > P(y_2|x)$

$y_2$ otherwise

**Note:**

Prior and likelihood are the main factors determining the posterior probability the evidence can be considered as scaling.

# Error Analysis

The probability of error is

$$
\begin{aligned}
P(error|x) &= P(y_1|x) \text{ if we decide } y_2 \\
&= P(y_2|x) \text{ if we decide } y_1
\end{aligned}
$$

The overall probability of error is

$$P(error) = \int_{-\infty}^{+\infty} P(error, x)\mathrm{d}x \quad = \int_{-\infty}^{+\infty} P(error|x)P(x)\mathrm{d}x$$

The Bayes decision rule says

$$
\begin{aligned}
&y_1 \text{ if } P(y_1|x) \; > \; P(y_2|x) \\
&y_2 \text{ otherwise}
\end{aligned}
$$

So, it minimizes $P(error|x)$. Hence $P(error)$ is also minimized

# Bayesian Decision Theory: A General Setting

| | | |
|---|---|---|
| $\{y_1, y_2, \ldots, y_c\}$ | : | a finite set of classes |
| $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ | : | a finite set of actions |
| $\lambda(\alpha_i \vert y_j)$ , $i = 1, 2, \ldots, a$ and $j = 1, 2, \ldots, c$ | : | denotes a loss function that describes loss for taking action $\alpha_i$ when the of the $x$ value is $y_i$ |
| $x \in \mathbb{R}^D$ | : | is a feature vector which is an instance of random vectors |
| $P(x \vert y_j)$, $j = 1, 2, \ldots, c$ | : | class conditional probability density function or likelihood |
| $P(y_j)$, $j = 1, 2, \ldots, c$ | : | prior probabilities |

- Posterior probabilities $P(y_i|x)$ $j = 1, 2, \ldots, c$ can be calculated using the Bayes formula $P(y_i|x) = \frac{P(x|y_i)P(y_i)}{P(x)}$
- where the evidence $P(x) = \sum_j^c P(x|y_j)P(y_j)$

# Bayesian Decision Rule as Risk Minimization

Suppose given $x \in \mathbb{R}^D$, we take action $\alpha_i$, then the expected loss associated with taking action $\alpha_i$ is

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

This is called the conditional risk. In continuous form overall risk is

$$R = \int_{x \in \mathbb{R}^D} R(\alpha(x)|x)P(x)\mathrm{d}x$$

# Bayesian Decision Rule as Risk Minimization

**Aim**: Find the decision rule that minimizes the overall risk $R$.

- The minimum risk is called the Bayes risk
- Suppose $\alpha^*(x) = \underset{\alpha(x)=\{\alpha_1,\alpha_2,\ldots,\alpha_a\}}{\arg\min} R(\alpha_i|x)$
- Then
$$R^* = \int_{x\in\mathbb{R}^D} R(\alpha^*(x)|x)P(x)\mathrm{d}x$$
is the minimum risk.

# Two Class Classification and Likelihood Ratio

- Let action $\alpha_i$ denotes deciding that true class label is $y_1$, $\alpha_2$ denotes deciding that true class is $y_2$
- Let $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ for $i = 1, 2$ and $j = 1, 2$, denotes the loss incurred when the decision is $\alpha_i$ but true class is $\omega_j$
- The conditional risk for any observation $x \in \mathbb{R}^d$ is

$$R(\alpha_1|x) = \lambda_{11}P(y_1|x) + \lambda_{12}P(y_2|x)$$
$$R(\alpha_2|x) = \lambda_{21}P(y_1|x) + \lambda_{22}P(y_2|x)$$

- Decision rule is

$$y_1 \text{ if } R(\alpha_1|x) \; < \; R(\alpha_2|x)$$
$$y_2 \text{ otherwise}$$

- Here we are taking decision based on the risk not by minimum posterior probabilities.

# Two Class Classification and Likelihood Ratio (contd...)

$$
\begin{aligned}
R(\alpha_1|x) &< R(\alpha_2|x) \\
\lambda_{11}P(\omega_1|x) + \lambda_{12}P(\omega_2|x) &< \lambda_{21}P(\omega_1|x) + \lambda_{22}P(\omega_2|x)
\end{aligned}
$$

- We have $\lambda_{21} = \lambda(\alpha_2|\omega_1)$ loss occurred for being wrong
- We have $\lambda_{11} = \lambda(\alpha_1|\omega_1)$ loss occurred for being right
- Similarly $\lambda_{12}$ and $\lambda_{22}$
- It is sensible to assume $\lambda_{21} > \lambda_{11}$ and $\lambda_{12} > \lambda_{22}$ as risk in being wrong is greater than for being right.
- So, $\lambda_{21} - \lambda_{11} > 0$ and $\lambda_{12} - \lambda_{22} > 0$
- Now by minimum risk strategy we decide $\omega_1$ if $(\lambda_{21} - \lambda_{11})P(\omega_1|x) > (\lambda_{12} - \lambda_{22})P(\omega_2|x)$ else $\omega_2$.

Now using Bayes theorem we write the previous strategy in terms of prior and likelihood as given below.

$$
\begin{aligned}
&(\lambda_{21} - \lambda_{11})P(y_1)P(x|y_1) > (\lambda_{12} - \lambda_{22})P(y_2)P(x|y_2) \\
&\implies \frac{P(x|y_1)}{P(x|y_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(y_2)}{P(y_1)} \\
&\implies \text{likelihood ratio} > \text{quantity independent of } x
\end{aligned}
$$

$\implies \psi(x) > c$, where $\psi(x) = \frac{P(x|y_1)}{P(x|y_2)}$

# Two Class Classification and Likelihood Ratio: Summary

- Bayes rule can be interpreted as deciding $y_1$ if the likelihood ratio exceeds a threshold value that is independent of $x$.
- Assumption is that we know the class conditional densities.
- In practical setting we learn likelihood from the training dataset. That is the threshold $c$ act as prior and $\psi(x)$ act as classifier whose parameters are to be learned from the data.

## Classification with 0-1 loss

- $\{y_1, y_2, ...., y_c\}$ a finite set of classes
- $\{\alpha_1, \alpha_2, ...., \alpha_c\}$ a finite set of actions corresponding to $\{y_1, y_2, ...., y_c\}$
- 0-1 loss is define as

$$\begin{aligned} \lambda(\alpha_i|y_j) &= 0 \; if \; i = j \\ &= 1 \; if \; i \neq j \\ i, j &= 1, 2, ..c \end{aligned}$$

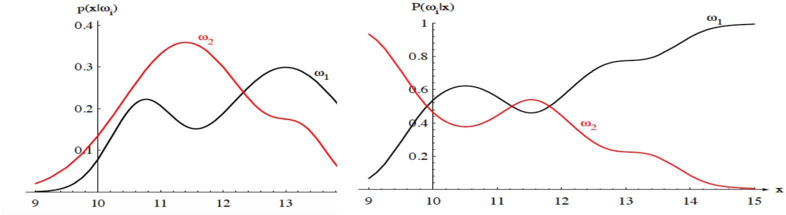
This assigns no loss to a correct decision and assigns unit loss to wrong decision. Now conditional risk

$$R(\alpha_i|x) = \sum_{j}^{c} \lambda(\alpha)i|y_j)P(y_j|x) = \sum_{j \neq i} P(y_j|x) = 1 - P(y_i|x)$$

$\implies$ If we decide on $y_i$ if $P(y_j|x)$ is maximum

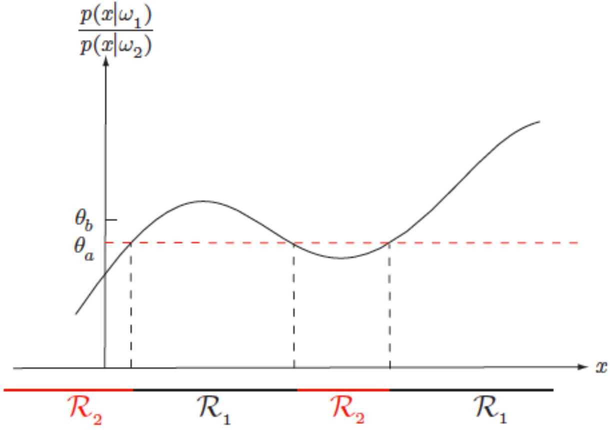
$\implies R(\alpha_i|x)$ is minimum $\implies R(x)$ is minimum

# Bayes rule in action




***(a)*** *Likelihood* ***(b)*** *Posterior*

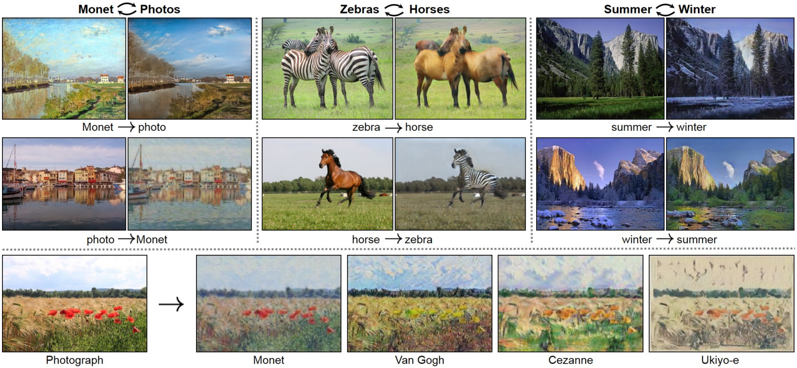
# Bayes rule in action




*Likelihood Ratio and threshold for decision boundary*

# Applications of Machine Learning

# Machine Learning 101 - Building the hype!




*CycleGAN: Image to Image Translation*[1]

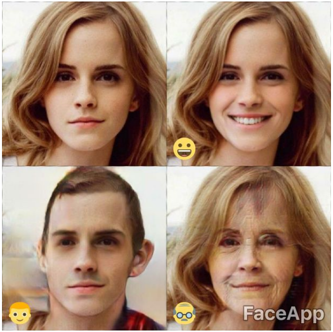
- Using video games to train autonomous driving systems
- More realistic image filtering in smartphone cameras etc.

[1] Image Source: `https://junyanz.github.io/CycleGAN/`

*Colorizing a Grayscale Image*[2]

- Converting all old movies into their colored version
- Restoring old paintings etc.

[2] Image Source: https://github.com/ImagingLab/Colorizing-with-GANs




*Neural FaceApp*[3]

- Victim identification during police investigations
- Smartphone filters etc.

[3] Image Source: Google

| Original Sentence | Flipped Sentiment |
|---|---|
| the film is strictly routine ! | the film is full of imagination. |
| after watching this movie, I felt that disappointed. | after seeing this film, I'm a fan. |
| the acting is uniformly bad either. | the performances are uniformly good. |
| this is just awful. | this is pure genius. |

Flipping sentiment of a sentence[4]

- De-radicalizing posts on Facebook
- Removing offensive sentences from movie captions

[4] Source: Toward Controlled Generation of Text

# Machine Learning 101 - Building the hype!

| **Dialogue** | **Dialogue** |
|---|---|
| SYSTEM: Hi. This is an Alexa Prize socialbot.<br>USER: Hi<br>SYSTEM: Do you have pets?<br>USER: Yes, I have two rabbits. | ...<br>USER: Let's talk about movies.<br>SYSTEM: What's your favorite movie?<br>USER: star wars movie |
| **Candidate Responses** | **Candidate Responses** |
| BoWFactGenerator: Did you know that male rabbits are called bucks, females are does?<br>GRUQuestionGenerator:<br>How about the American family of it?<br>Initiatorbot:<br>Do you believe in love at first sight?<br>LSTMClassifierMSMarco:<br>Mar 23, 2017 Two Parts: Learning What to Feed Your Rabbit Learning How to ...<br>DualEncoderRedditNews: bye doggie<br>**Alicebot:**<br>**Hurrah! Two is a good number of rabbits.**<br>Elizabot: OK, but can you elaborate a bit?<br>VHREDRedditMovies:<br>Hmm. Nope, never heard of it.<br>VHREDSubtitles: Would you like to see them?<br>BoWWashingtonPost:<br>I do but I call people with dogs crazy also. | BoWFactGenerator: Here's an interesting fact. The international telephone dialing code for Antarctica is 672.<br>GRUQuestionGenerator: how about your New York City?<br>Initiatorbot: How was your day?<br>LSTMClassifierMSMarco:<br>A third Anthology film will be released in 2020.<br>DualEncoderRedditNews:<br>Now whisper it to me , one more time bby.<br>Alicebot: What did you like about the robots in that movie?<br>Elizabot: How does that make you feel?<br>**Evi: Star Wars movie a movie in the Star Wars series.**<br>VHREDRedditMovies:<br>Oh please. Please. Pleeeease. Let this happen.<br>VHREDSubtitles: What?<br>BoWWashingtonPost:<br>A much more enjoyable feature than last year's old-timer's convention masquerading as a star wars movie. |

*Chatbots*[5]

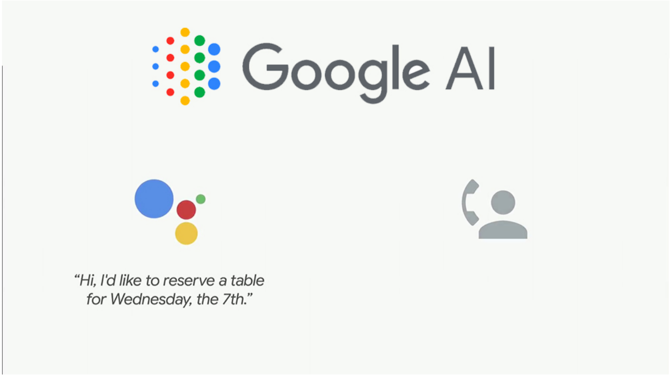
- In personal assistants like Siri, Google Assistant etc.
- Challenges include sustaining a long range conversation etc.

[5] Image Source: A Deep Reinforcement Learning Chatbot


*Visual Question Answering*[6]

- Transcribing videos to generate documentation of a procedure
- Helping blind people in sensing the world around them

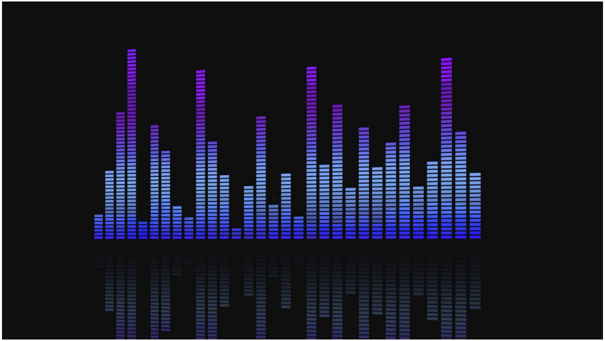
[6] Image Source: Making V in VQA Matter




*Speech Generation*[7]

- Talking in a real world setting
- Personal assistants

[7] Image Source: Google



*Generating Music*[8]

- Conditionally generating music
- Can we replace the monotonous music at customer cares and personalize it to users?

[8] Image Source: Google

# Machine Learning 101 - Building the hype!

- Find topics from billions of documents in completely unsupervised way
- Used for improving search results, categorizing documents, finding trends in literature etc.
- The most commonly used algorithm (LDA) is efficient enough to run on a single laptop

| Theme | Description | Top words |
|---|---|---|
| State bans | State level regulations on abortion | ban, state, govt, bill, ohio |
| Women's rights | Abortion as women's fundamental right | women's, rights, pills, reproductive, healthcare |
| Religious views | Church's stance on abortion | jesus, religion, bible, god, faith |
| Abortion is murder | Perceiving abortion as an act of killing | kill, murder, wrong, life, baby |
| Planned Parenthood | organization for reproductive health services | planned, parenthood, defund, pp, clinics |

*Topic Modeling*[9]

**Countless other Applications:**

- **Biology and Medicine:**
  - Protein interaction prediction
  - Automated drug discovery
  - Predicting diseases faster than human experts etc.
- **Security:**
  - Applications like face recognition
  - Detecting fraudulent transactions
  - Automated video surveillance etc.
- **Social Sciences**
  - Spreading ideas in a social network
  - Friend recommendations
  - Analyzing large scale surveys etc.

# Machine Learning 101 - Building the hype!

- **Information Extraction**
  - Web search
  - Question answering
  - Knowledge graph mining etc.
- **Economics and Finance**
  - Algorithmic Trading
  - Analyzing purchase patterns and market analysis
  - e-commerce applications like product recommendations etc.
- **Others**
  - Automated theorem proving
  - Robotics
  - Advertising
  - And many more. . .

# Summary

- What involves developing machine learning algorithms?
- Classification using Bayes rule: Incorporating prior knowledge
- Yes! Machine learning is very exiting field and it has many applications

**References:**

- Chapter 2, Pattern Classification by Duda, Hart and Stork