

1 Overview

- Week 1-3: Classical AI, search algorithms
 1. Uninformed search
 2. Local search: hill climbing
 3. Informaed search: A*
 4. Adversarial search Minimax
- Week 4-7: Classical ML
 1. Decision trees
 2. Linear/Logistic regression
 3. Kernels and support vector machines
 4. "Classical" unsuperivese learning
- Week 10-12: Modern ML
 1. Neural networks
 2. Deep learning
 3. Sequential data
- Week 13: Misc.

2 AI: Computers Trying to Behave Like Humans

- **PEAS Framework:**
 - **Performance measure:** define “goodness” of a solution
 - **Environment:** define what the agent can and cannot do
 - **Actuators:** outputs
 - **Sensors:** inputs
- Agent function is sufficient.
- Common agent structures (to define an AI agent):
 - Reflex
 - Goal-based
 - Utility-based
 - Learning
 - (Others possible; can mix and match!)
- Exploration vs exploitation

3 Problem Statement

fully observable \wedge deterministic \wedge static \wedge discrete \implies only need to observe once

To solve a prob using search:

- A goal or a set of goals
- a model of the environment
- a search algorithm

goal formulation \rightarrow problem formulation \rightarrow search \rightarrow execute

1. goal formulation

2. problem formulation, eg. path finding

- states: nodes representation invariant:: abstract states should correspond to concrete states
- initial state: starting node
- goal states/test: dest node
Goal test: define the goal using a function *is_goal*
- actions: move along an edge :: $|actions(state)| \leq branching_factor$
- transition model: $f(curr_state, action) \implies next_state$
- action cost function: see edges

3. Important facts:

- Representation Invariant: ensure that the abstract states correspond to concrete states
- Goal Test: Goal defined via a function *is_goal*
- Action: a set of $action(state)$, $|actions(state)| \leq branching_factor$
- Transition model: $f(curr_state, action) \implies next_state$

Search

Uninformed search

No information that could guide the search: no clue how good a state is

```
create frontier
// create visited // with visited memory
insert Node(initial_state) to frontier
while frontier is not empty:
    node = frontier.pop()
    if node.state is goal:
        return solution
// if node.state in visited: // with visited memory
//     continue
// visited.add(state)
for action in actions(node.state):
    next_state = transition(node.state, action)
    frontier.add(Node(next_state))
return failure
```

Different subvariant of tree search uses different DS for the frontier.

Search Type	Data Structure for Frontier
BFS	Queue
DFS	Stack
UCS (Uniform-cost Search)	Priority Queue

Depth limited search

limit the search to depth l

backtrack when the limit is hit.

time complexity: exponential to search depth

space complexity: size of the frontier

```
create frontier
tier = 0
insert Node(initial_state) to frontier
while (!empty(frontier)) && (tier <= limit):
    node = frontier.pop()
    tier++
    if node.state is goal:
        return solution
    for action in actions(node.state):
        next_state = transition(node.state, action)
        frontier.add(Node(next_state))
return failure
```

Iterative deepening search

search with depth from 0 to ∞

return soln when found. Both complete

```
create frontier
tier = 0
insert Node(initial_state) to frontier
while (!empty(frontier)) && (tier <= limit):
    node = frontier.pop()
    tier++
    if node.state is goal:
        return solution
    for action in actions(node.state):
        next_state = transition(node.state, action)
        frontier.add(Node(next_state))
return failure
```

Summary

Name	Time Complexity*	Space Complexity*	Complete?	Optimal?
Breadth-first Search	Exponential	Exponential	Yes	Yes
Uniform-cost Search	Exponential	Exponential	Yes	Yes
Depth-first Search	Exponential	Polynomial	No#	No
Depth-limited Search	Exponential	Polynomial**	No**	No**
Iterative Deepening Search	Exponential	Exponential**	Yes	Yes

Not complete if not tracking visited nodes, search may stuck in loop before visiting all nodes.

* In terms of some notion of depth/tier

** If used with DFS

4 Local Search

Systematic search: typically complete and optimal under certain constraints. However intractable sometimes
Local search: typically incomplete and suboptimal, but has anytime property, ie. longer time -> better solution. Able to provide good enough solution under reasonable amount of time.

4.1

1. Start at random position in the state space
2. iteratively move from a state to another neighbouring state via perturbation or construction
3. solution is the final state

State space: all possible configuration (1)

Search space: a subset of state space that will be explored (2)

4.1.1 Perturbation search

- Search space: complete candidate solutions
- search step: modification of one or more solution

For example: swap a path with another path

4.1.2 Constructive search

- partial candidate solution
- extension of one or more solution

For example: path finding

4.2

goal formulation -> problem formulation -> search -> execute

1. goal formulation
2. problem formulation, eg. path finding
 - states: nodes representation invariant:: abstract states MAYNOT directly correspond to concrete states
 - initial state: starting node, a candidate solution
 - goal states/test: dest node [optional]
Goal test: define the goal using a function $is_goal\ f(curr_state, action) \implies next_state$
 - Successor function: a function that generates neighbouring states by applying modifications from the current state. This defines the local search space

4.3 Evaluation function

A math function that assess the quality or desirability of the solution.

Some solutions may be unacceptable but there are some less bad than the others.

4.4 Hill Climbing/ Greedy Local Search

```
curr_state = init_state
while 1:
    best_succ = best(successor(curr_state))
    if (eval(best) <= eval(curr_state))
        return curr_state
    curr_state = best_succ
```

4.5 State space landscape

- Global max:
- Local max:
- shoulder:

5 Adversarial search

5.1 Classical adversarial games

- Fully observable
- Deterministic
- discrete
- No infinite run
- 2-player zero-sum
- turn taking

terms

- Player: agent
- Turn:
- Move
- End state
- winning conditon
-

5.2 Problem formulation in adversarial search

- states: nodes representation invariant:: abstract states MAYNOT directly correspond to concrete states
- initial state: starting node, a candidate solution
- Terminal State: state the outcome of the game when it terminates
- Utility function: output the value of a state from the perspetive of our agent

5.3 Minimax

In the view of A, A try to maximize the outcome of the game, B will try to minimize A's outcome, as the gam is zero sum

- $\text{expand}(\text{state}) \Rightarrow [a]$

```
max_value(state):
    if is_terminal(state): return utility(state)
    v = -∞
    for next_state in expand(state):
        v = max(v, min value for player A in next_state)
    return v

min_value(state):
    if is_terminal(state): return utility(state)
    v = ∞
    for next_state in expand(state):
        v = min(v, max value for player A in next_state)
    return v

minimax(state):
    v = max_value(state)
    return action in expand(state) with value v
```

5.4 Alpha-beta prunning

From the viewpoint of the MAX player:

- α : the value of the best choice for the MAX player so far
- β : the value of the best choice for the MIN player so far

An optimized version of the Minimax algorithm using pruning:

```
max_value(state, α, β):
    if is_terminal(state): return utility(state)
    v = -∞
    for next_state in expand(state):
        v = max(v, min(v, α, β))
    return v

min_value(state, α, β):
    if is_terminal(state): return utility(state)
    v = ∞
    for next_state in expand(state):
        v = min(v, max(v, α, β))
    return v

alpha-beta search(state):
    v = max_value(state, -∞, ∞) // initialized α to be -∞, β to ∞
    return action in expand(state) with value v
```

6 Learning agent

For problems that the function are difficult to specify, solutions re intractable to compute in general. Typically episodic,

$$DL \subset ML \subset AL$$

6.1 supervised

Learn the mapping input, feedback given \rightarrow output, given a dataset, it minimizes the difference between the prediction and the provided correct ans using a leanign algorithm e.g. image ientification

1. Train phase: try minimizing the diff between pred and correct ans given using the training set, resulting in a trainign agen function, known as the model/hypothesis
2. Testing/evaluation phase: using a test set to measure the performance of the model. The performance on unseen data measures the generalization of the model

Task

- Classification: to predict discrete labels or catagories on the input features
- Regression: predict continuous numetical value based on input features

Dataset

$$D = \bigcup_{[1,n]} \{(x^{(i)}, y^{(i)})\}$$

True data generation function

$$y = f^*(x) + \epsilon$$

where $f^*(x)$ is true but unknow, which generates the label from the input features; ϵ is some noise or error term, which account for the randomness or imperfection in the date generating process.
the goal is to find a function that best approximately $f^*(x)$

Hypothesis class

THE set of models or functions that maps from inputs to outputs $h : X \implies Y$ that can be learned by a learing algorithm. Each element of the hypo clas $h \in \mathcal{H}$

LEarnign algorithm

$$A(D_{train}, H_{hypo}) = h(x), h \in \mathcal{H} \approx f^*(x)$$

PERformance measure

$$h(x) \approx f^*(x)$$

$$PM(D_{test}, h \in H_{hypo}) \mapsto$$

Try the hypothesis h on a new set of examples (test data)

Regression: error

$$\text{Absolute error} = |\hat{y} - y| \quad (3)$$

$$\text{Squared error} = (\hat{y} - y)^2 \quad (4)$$

Mean squared error

$$MSE = \frac{1}{N} \sum_{i=0}^N (\hat{y}^{(i)} - y^{(i)})^2$$

Mean absolute error

$$MAE = \frac{1}{N} \sum_{i=0}^N |\hat{y}^{(i)} - y^{(i)}|$$

where

$$\hat{y}^{(i)} = h(x^{(i)})$$

Accuracy:

$$A = \frac{1}{N} \sum_{i=1}^n \mathbb{1}_{\hat{y}^{(i)} = y^{(i)}}$$

confusion matrix

True positive, false positive, false negative, true negative (2, 1, 3, 4 quadrant)

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

maximize if FP is costly

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

maximize recall if FN is dangerous

$$F1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$

6.2 Decision Tree

Greedy, top-down, recursive, use

```
DTL(examples, attributes, default):
    if (examples == ∅): return default
    if (∀e ∈ examples, e has the same classification c): return c
    if (attributes == ∅): return mode(examples)

    best = choose_attribute(attributes, examples)
    tree = new decision tree with root test best
    for vi of best do:
        examples_i = {e | e ∈ examples, e.best = vi}
        subtree = DTL(examples_i, attributes \ best, mode(examples_i))
        tree.add(vi: subtree)
```

$$f : [\text{attribute vector}] \mapsto \text{Boolean}$$

Basically nested if-else

Expressiveness

Decision trees can express any function of the input attributes. Trivially, Consistent training set \rightarrow Consistent decision tree, but unlikely to generalize to new examples

Each row in the truth table is represented as a path in the decision tree

Size of the hypothesis class

For n boolean attributes, there are n boolean functions, n distinct truth tables each with 2^n rows $\rightarrow 2^{2^n}$ decision trees

Informativeness

Ideally we want to select an attribute that splits the examples into all positive or all negative.

Entropy: The higher the more random, thus less informative

$$I(P(v_1) \dots P(v_k)) = - \sum_{i=1}^k P(v_i) \log_2 P(v_i)$$

For data set contains boolean outputs,

$$I(P(+), P(-)) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$

where $0 \leq \mathbb{R}_I \leq 1$. However for non-binary variables the entropy can be greater than 1

Information gain

Information gain = entropy of this node - entropy of children nodes

$$IG(A) = I\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - remainder(A)$$

$$remainder(A) = \sum_{i=1}^v \frac{p_i + n_i}{p+n} I\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

Decision tree pruning

- By sample size
- by max search depth

6.3 unsupervised

Find pattern. No feedback given. e.g, group images by cluster

6.4 Reinforcement

Trial and error, reward given based on observation and action. eg. chess

7 Linear regression

7.1 LR

- Data: N data points: $[(feature_vector, target)]$

Regression: given $x \in \mathbb{R}^d$ and no target, find a function that predicts the target $y \in \mathbb{R}$
the function:

$$f : \mathbb{R}^d \mapsto \mathbb{R}$$

Linear Model:

$$h_w(x) = \sum_{i=1}^d w_i x_i = v_w^T x$$

Where w_i are the parameters or weights and $x_0 = 1$ is a dummy variable (always 1), w_0 is the bias. when $d = 1$, the model is linear

Feature Transformation

Modify the original features of a dataset to make them more suitable for modeling.

- Feature engineering
 - Polynomial features: $z = x^k$, k is the polynomial degree.
 - log feature: $z = \log(x)$
 - Exp. feature: $z = e^x$
- Feature scale
 - Min-max scaling: $z_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$, scales to $[0, 1]$
 - standardization: $z_i = \frac{x_i - \mu_i}{\sigma_i}$, transformed data has mean of 0 and SD of 1
 - robust scaling (not in syl)
- Feature encoding (not in syl)

Measuring Fit

Loss function (MSE of the for N samples):

$$J_{MSE}(w) = \frac{1}{2N} \sum_{i=1}^N (h_w(x^{(i)}) - y^{(i)})^2 \quad (5)$$

Since polynomial regression can be converted into linear regression, by theorem 7.2, the MSE loss function on linear/poly nomial regression is convex

7.2 Normal Equation

Minimize a function: partial derivative

$$f(w) \mapsto \frac{\partial(f(w))}{\partial(w_i)} \quad \text{where} \quad f(w) = h_w(x^{(i)})$$

1. Take the partial derivative of the linear model

$$\frac{\partial(h_w(x^{(i)}))}{\partial w_j} = \frac{\partial(w^T x^{(i)})}{\partial w_j} = \frac{\partial(\sum_{i=1}^N w_i x_i)}{\partial w_j}$$

2. Take the par. deriv. of each term in the sum, with respect to w_0 and w_1

$$\frac{\partial((h_w(x^{(i)}) - y)^2)}{\partial(w_j)} = 2(h_w(x^{(i)}) - y)x_j$$

Where

$$h_w(x^{(i)}) = w_0 + w_1(x^{(i)})$$

$$\frac{\partial}{\partial w_0} \left[\frac{1}{2N} \sum_{i=1}^N \left((w_0 + w_1 x^{(i)}) - y^{(i)} \right)^2 \right] = \frac{1}{N} \sum_{i=1}^N ((w_0 + w_1 x^{(i)}) - y^{(i)}) \quad (x_0 = 1)$$

$$\frac{\partial}{\partial w_1} \left[\frac{1}{2N} \sum_{i=1}^N \left((w_0 + w_1 x^{(i)}) - y^{(i)} \right)^2 \right] = \frac{1}{N} \sum_{i=1}^N ((w_0 + w_1 x^{(i)}) - y^{(i)}) (x^{(i)})$$

In generalized form, where $N > 1$,

$$\frac{\partial}{\partial w_N} \left[\frac{1}{2N} \sum_{i=1}^N \left(h_w(x^{(i)}) - y^{(i)} \right)^2 \right] = \frac{1}{N} \sum_{i=1}^N (h_w(x^{(i)}) - y^{(i)}) (x_N^{(i)})$$

Normal Equation

find w that minimizes J_{MSE} :

$$\frac{\partial(J_{MSE})}{\partial(w_j)} = \frac{1}{N} \sum_{i=1}^N (w^T x^{(i)} - y^{(i)}) x_j^{(i)} = 0$$

Where $w = \begin{bmatrix} w_0 \\ w_1 \\ \dots \\ w_N \end{bmatrix}$, $x = \begin{bmatrix} x_0 \\ x_1 \\ \dots \\ x_N \end{bmatrix}$, $w^T x$ is just the linear combination of feature vector w on x

$$X^T(Xw - Y) = 0$$

Suppose invertible

$$w = (X^T X)^{-1} X^T Y$$

7.3 Gradient Descent

- GD Algorithm
- Variant: mini-batch, stochastic
- Problems and Solutions

Computing the partial derivative to find the minimum is costly, use local search to find a local min

$$w_j \leftarrow w_j - \gamma \frac{\partial(J(w_0, w_1 \dots))}{\partial(w_j)}$$

Where γ is the learning rate.

Repeat until termination criterion is satisfied

Convexity

Theorem 7.1. *A convex function has one singular global minimum.*

Theorem 7.2. *MSE loss function is convex for linear regression*

Variants

- Batch: take all training samples
- Mini-batch: random subset
- Stochastic: select one random sample

Linear regression with Regularization

Knowing the hypothesis

$$h_w(x) : w^T x$$

Regularized cost function

$$J(w) = \frac{1}{2m} \left[\sum_{i=1}^m (h_w(x^{(i)}) - y(i)) + \lambda \sum_{i=1}^n w_i^2 \right]$$

Where

$\sum_{i=1}^m (h_w(x^{(i)}) - y(i))$ fits the data "well", where $\lambda \sum_{i=1}^n w_i^2$ is the regularization parameter, that avoid "over-fitting"

$$w = (X^T X)^{-1} X^T Y$$

$$w = (X^T X + \lambda \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix})^{-1} X^T Y$$

As $w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{bmatrix}$, here in the regularized cost function, m indicates the size of the training set, while n

indicates the size of the feature vector large $\lambda \rightarrow$ underfitting; $\lambda = 0 \rightarrow$ overfitting

8 Logistic Regression

Classification Task

$$x \in \mathbb{R}^d \mapsto \text{void}$$

Find a function that

$$y \mapsto \{0, 1\}, \text{ for each } x$$

Sigmoid

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$
$$\sigma'(x) = \frac{-(1 + e^{-x})'}{(1 + e^{-x})^2} = \frac{e^{-x}}{1 + 2e^{-x} + e^{-2x}}$$

$$\sigma(f(x))$$

$$\sigma(x^2) = \frac{1}{1 + e^{-x^2}}$$

Decision Boundary

A surface that separates the different classes in the feature space

The intersection between the hypothesis function and the decision threshold

Binary Cross Entropy

$$BCE(y, \hat{y}) = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y})$$

$$J_{BCE} = \frac{1}{N} \sum BCE_i$$

$\sigma(x) = \frac{1}{1+e^{-x}}$ is Not convex, but $-\log(\sigma(x))$ is convex
Hypothesis function

$$h_w(x) = \sigma(w_0, w_1x_1 + w_2x_2)$$

Weight Update

$$w_j \leftarrow w_j - \gamma \frac{\partial J_{BCE}(w_0, w_1, \dots)}{\partial w_j}$$

8.1 Logistic regression cost function

$$\begin{cases} -\log(h_w(x)), & \text{if } y = 1 \\ -\log(1 - h_w(x)), & \text{if } y = 0 \end{cases} \quad (6)$$

Multiclass classification

One-to-one

$$\{0, 1, \dots, C\} \times \{0, 1, \dots, C\} \rightarrow (C^2 - C) \text{ BCE}$$

One-vs-Rest

$$\{0, 1, \dots, C\}.map(c \rightarrow \{0, 1, \dots, C\} \setminus \{c\})$$

Generalization

Data Quality

Data Quantity

Model Complexity