Chapter 2 Inductive learning

1.Welche Ziele verfolgen Induktion und Deduktion? Wie erreichen Sie diese Ziele?

Induction: From specific observation derive general principle

Deduction: From general principle derive specific logical conclusion.

2.Definieren Sie Konsistenz und Vollständigkeit im Hypothesenraum.

Consistence: No negative examples are classified as positive.

Complete: All positive examples are classified as positive.

3. Version space

Version space is a subset of Hypothesis space, where all hypotheses are consistent and complete.

4. Inductive learning hypothesis

Any hepythesis, that approximates target function well enough over a suffficently large training set, will also approximate the target function well over unknown examples.

5. Inductive bias

Set of assumptions or prior knowledge that learning system incorporates to generalize from data. Some hypotheses are preferred than other hypotheses.

2 Kinds: hypothesis I that maximizes the distance between 2 classes is preferred

The simpler hypothesis is preferred.

6.Ausgehend von der lerntheoretischen Abschätzung des realen Fehlers, von welchen drei Kriterien hängt der Lernerfolg einer Lernmaschine ab?

Complexity of model, size and quality of training data, optimization and regularization.

Chapter 3 Learning Theory

1.Warum wird in der Praxis der empirische Fehler anstelle des realen Fehlers berechnet?

Normally the real error can not be calculated in finite time, since the amount of calculation is too large.

2.Nennen Sie jeweils einen Vor- und einen Nachteil von Bagging.

Vorteil: Improving the quality of model.

Nachteil: Increasing the difficulty and amount of computation of training.

3.Wie ist die Vapnik-Chervonenkis (VC) Dimension für lineare Klassifikation definiert?

The VC Dimension of H is equal to the maximum number of datapoints which can be arbitrarily separated from H.

4. Reason and solution of overfitting

Reason: capacity of model is too high; the model is trained for too many iterations.

Solution: decrease the capacity; early stopping; find optimal hypothesis

5. Overfitting formal

There exists other hypothesis in hypothesis space, which fits less well on training data but performs better over the entire distribution of instances.

h overfits $\Rightarrow \exists h' \in H$, such that given D_{Tr} and D_V

$L_{D_{Tr}}(h') > L_{D_{Tr}}(h)$ and $L_{D_V}(h') < L_{D_V}(h)$

6.Verfahren des maschinellen Lernens lassen sich in unterschiedliche Kriterien einordnen. Nennen Sie zwei davon und geben Sie deren Ausprägungen oder Abstufungen an

Unsupervised learning, Superwised learning

7.Nennen Sie zwei Probleme, die bezüglich der Ausartung der Fehlerflächen beim Gradientenabstieg auftreten können. Geben Sie zwei Methoden an, mit denen diese Probleme jeweils vermieden werden können.

Stuck in local minimum: cycling learning rate

Saddle point: momentum, Adam

8.Wie unterscheidet sich Stochastic Gradient Descent (bzw. Pattern Learning) vom „echten“ Gradientenabstieg? Was sind die jeweiligen Vorteile der beiden Verfahren?

SGD: Gradients are updated after the randomly sampled minibatch is processed

GD: Gradients are updated after the whole dataset is processed.

Vorteil: SGD: quicker convergency

GD: stable convergency for convex function

9.Beschreiben Sie kurz den theoretischen Nutzen in der Anwendung.

Measure of complexity of learning data.

Measure of capacity of learning system.

Chapter 4 Decision Tree

1.Welches Problem lösen Random Forests im Vergleich zu Bagging, wenn Entscheidungsbäume als Modell verwendet werden? Durch welche Modifikation des Bagging Algorithmus wird dies erreicht?

Problem: The Models generated by Bagging are highly correlated.

Modification: At each split we randomly choose a subset of all attributes as possible candidate attributes for splitting.

2. Reduce Overfitting for ID3

Maximum depth: set a maximum depth for the decision tree, if the maximum depth is reached, stop splitting

Minimum samples: when the number of samples of a node is fewer than minimum samples, stop splitting

Early stopping: stop training when validation error increases

Pruning: Cut the non-critical branches to reduce complexity

Bagging and random Forrest.

3.Bagging – Pros & Cons

Pros:

Uncertainty, Out-of-bag Error(unsampled data as validation set), increase variance,parallelizable.

Cons:

Models are highly correlated, computationally expensive, loss of interpretablity.

4. Was ist das Optimierungsziel beim Lernen von Entscheidungsbäumen bzgl. der Klassenverteilung? Welches Maß wird hierfür herangezogen? Wie wirkt sich dies auf die tiefe des gelernten Baums aus?

Ziel: Classify the data into corresponding classes as quick as possible

Mass: At each splitting select the attribute with minimum entropy maximum information gain

Auswirkung: tree with little depth

5.Nennen Sie vier positive Eigenschaften eines Random Forests (verglichen mit einem einfachen Entscheidungsbaum).

Reduce overfitting, improved accuracy, handling unbalanced data, versatility in handling different data types

6. Nennen Sie zwei Vorteile von Entscheidungsbäumen gegenüber neuronalen Netzen.

1.Decision tree is non-parametric

2.No preprocess of data is needed

3. Better interpretability.

Chapter 5 and 6 Neural Networks

1.Optimierungsmethoden zweiter Ordnung sollten den Loss besser minimieren. Dennoch werden sie bei neuronalen Netzen nur sehr selten verwendet. Wieso?

Hessian contains O(n^2) values for n parameters, which requires a lot of memory.

2.Nennen Sie zusätzlich eine Optimierungsmethode erster Ordnung, die eine Optimierungsmethode zweiter Ordnung approximiert.

Momentum and Adam.

3.Wie können sich zu groß oder zu klein gewählte Lernraten auf den Trainingsverlauf des Modells auswirken?

Small learning rate: It takes too many steps to find the minimum and can be stuck in local minimum.

Large learning rate: It will overshot the minimum even result in divergence.

4.Welchen Vorteil können dynamischen Lernraten, wie die häufig verwendete Cosinus- Reduktion (Cosine Annealing), bieten?

At the beginning the learning rate is relatively large, which can reduce the loss quickly. At the end the learning rate is small, which can find the minimum accurately.

5.Inwiefern ändert sich das Gewichtsupdate wenn, statt einem Eingabevektor, ein Mini-batch mit mehreren Eingabvektoren für das Training verwendet wird?

One-Vector Input: The weights will be updated after the whole dataset is processed.

$w^{k+1} = w^k - \eta \cdot \frac{1}{|N|} \sum_{i=1}^{N} \nabla L(y_i, \hat{y}_i)$ with N: Dataset

Mini-batch input: The weight will be updated after the Mini-batch is processed.

$w^{k+1} = w^k - \eta \cdot \frac{1}{B} \sum_{i=1}^{B} \nabla L(y_i, \hat{y}_i)$ with B: Mini-Batch

6.Was gilt es im Zusammenhang mit dem Dropout zu beachten, wenn das Training abgeschlossen wurde und das Netzwerk in der Praxis eingesetzt werden soll?

In training process every neuron can be deactivated with probability p. But in practice use every neuron will be always activated. So the weight parameters in practice should multiply the probability p.

7. L1 and L2 regularization(weight decay)

$\tilde{L} = \hat{L} + \lambda R(\theta)$

original loss Regularization term

$\lambda \geq 0 \quad R(\theta) = \begin{cases} L1: \|w\|_1 = \sum w_i \\ L2: \frac{1}{2}\|w\|_2^2 = \frac{1}{2} w^T \cdot w \end{cases}$

Advantages of L1:

Unimportant features are set to 0(sparse result)

Disadvantages of L1:

Constant factor that does not scale with specific values of *wi*

8. Con and Pros of all kinds of activation function(P39-P43) Softmax

9.Was muss bei der Initialisierung der Gewichte eines Neuronalen Netzes beachtet werden?

Nennen und erläutern Sie kurz eine gängige Initialisierungsmethode.

The weights should be neither too large nor too small, otherwise exploding or vanishing gradient occur.

Xavier: Keep variance of activation function constant for each layer. Suitable for tanh and sigmoid.

Kaiming He: Suitable for ReLu, if ReLu halves the variance in each layer, then increase the variance manually by that factor.

10.Welchen Vorteil können zyklische Lernraten gegenüber monoton absinkenden Reduktionsverfahren haben?

The large learning rate at the beginning of every cycle can help to escape the local minima.

Chapter 7 CNNs

1.Wieso werden GPUs statt CPUs für das Training und Ausführen von CNNs verwendet?

GPU has parallel calculation ability and is optimized for matrix operation, which is relevant in CNNs.

2.Wofür werden entweder Pooling Layer oder Strided Conv. Layer verwendet? Nennen Sie für beide Verfahren jeweils ein Vorteil.

They are used to reduce resolution.

Vorteil:Pooling layer: No learnable parameters.

Strided convolutional layer: most Information from last layer will be remained

3. Wie wird das Vanishing-Gradient Problem bei ResNet gelöst? Erläutern Sie dies kurz.

In Resnet, each input has a skip-connection to the output. When the trained weights of the layer is not reasonable, then we skip this layer, use identity of input as output of this layer.

4.Welche Vorteile haben 3 × 3 Kernel gegenüber größeren z.B. 7 × 7 Kernel hinsichtlich des rezeptiven Felds?

To achieve receptive field of size 7*7 for ever single pixel, we need

{ one 7x7 Kernels ⇒ 7x7+1=50 parameters
  or
  three 3x3 Kernels ⇒ 3x3x3+3=30 parameters

3*3 kernel is preferred, because it has fewer parameters

5.Nennen Sie drei Methoden der Gewichtsinitialisierung für CNNs. (___/2P)

Was passiert, wenn alle Gewichte mit 1 initialisiert werden?

Different approaches: Xavier Kaiming he.

Transition learning

6.~~Bei Computer Vision~~ Aufgaben werden heutzutage keine herkömmlichen neuronalen Netze verwendet sondern CNNs. Nennen Sie dafür zwei Gründe

Less learning parameters and translation-invariance

7.Was ist der Zweck einer 1x1 Faltung, wie sie im Inception Modul vorkommt?

Reduce number of channels. Increase nonlinearity

Chapter 8 Support Vector Machine

1.Nennen Sie zwei Methoden, mit denen die SVM auch auf nicht linear separierbaren Daten anwendbar ist.

Soft margin method and Kernel-Trick method

2.Erläutern Sie kurz den Kernel-Trick und welcher Vorteil sich daraus ergibt.

Kernel-Trick transform the data space into a higher dimension, where the nonlinear problem can be solved linearly.

<xi,xj> — <φ(xi),φ(xj)> —K(xi,xj) with K: kernel function

Vorteil:Compared with calculating the complex nonlinear transformation ϕ( ) and the scalar product, kernel-trick saves a lot of computational operations.

3.Kernel functions

Dot-product kernel function, polynominal kernel function, radial-basis kernel function

4.Bezogen auf einen Support Vektor Klassifikator, beschreiben Sie das Problem welches gelöst wird, die Lösung die gefunden wird und Intuition für die Lösung.

Problem: binary classification problem

Ituition: Find a decision boundary(hyper plane), which maximaize the margin between 2 classes.

5.Nennen Sie jeweils zwei Vor- und Nachteile von Support Vector Machines.

Pros:optimal hyperplane leads to good result, rapid evaluation of processing high dimensional data

Cons: high requirements of memory and computation power, hard to find a optimal kernel

Chapter 9 Unsupervised Learning

1.Im DBSCAN Algorithmus werden Datenpunkte in folgende drei Klassen eingeteilt:

1. Kernpunkte 2. Erreichbare Punkte 3. Rauschen

Erklären Sie für jede Klasse wann ein Datenpunkt ihr zugeteilt wird.

Hinweis: Sie dürfen die Variablen ϵ und minPoints als gegeben ansehen.

Core sample: A Point is a core sample if there are at least minpts points within the distance ε.

Neighbor: A point is a neighbor, if there are not minpts points but at least one core sample within the distance ε.

Noise: There is no core samples within distance ε.

2. Advantages and disadvantages of DBSCAN

Advantages: Robust against noise; priori knowledge of number of clusters is not needed; It can cluster different data densities.

Disadvantages: Sometimes non-deterministic; Still depend on distance function; appropriate selection of ε and minpts can be difficult.

Chapter 12 13 Reinforcement learning

1. Was ist das primäre Ziel von Reinforcement Learning? Geben Sie eine kurze Beschreibung an.

The Agent interacts with environment to maximize the expected cumulative reward.

2.Was ist die Hauptbedingung, die ein Prozess erfüllen muss, um als Markov-Entscheidungsprozess bezeichnet zu werden?

$P(S_{t+1} | S_t) = P(S_{t+1} | S_1, S_2 \cdots S_t)$

Next state is only related to the current state. Current state includes the information of all past states.

3.Definieren Sie kurz den Begriff „modellfrei“ im Kontext von Reinforcement Learning.

In reinforcement learning, the model is often unknown for the agent. The agent needs to interact with the environment to gather experience samples. From the experience, agent learn a policy, value functions or a model.

4.Exploration and Exploitation

Exploration: extending the current knowledge by choosing the action, that appear to be suboptimal but may potentially lead to higher returns

Exploitation: Choose the best action based on current knowledge

Exploration and exploitation trade off: The agent should balance between exploration and exploitation so that all actions are sufficiently explored but also maximize the return.

A greedy policy can get stuck on a sub-optimal action, because it only use exploitation. (Solution:ϵ-greedy policy,greedy action:1-ϵ,random action:ϵ)

5. Durch welches Entscheidungsmodell lässt sich die Problemstellung beim Reinforcement Learning formal darstellen?
Welche vier Bestandteile werden für die Modellierung benötigt?
Markov decision process.
States, state transition probability function, actions, reward function

6. Gegeben ist $V\pi(s)$: Bestimmen Sie die Optimale Strategie formal. Außerdem, definieren Sie die rekursive Form der Bellmann Optimalitätsgleichung in Abhängigkeit von der optimalen Wertfunktion.

Optimal policy: if $V_{\pi^*}(s) \geq V_\pi(s)$, $\forall s \in S$, then π^* is optimal policy

equation:
$$\begin{cases} V_{\pi^*}(s) = \max_{a \in A}\left(r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V_\pi(s')\right) \\ q^*(s) = r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) \max_{a' \in A} q(s',a') \\ V(s) = E(R_{t+1} + \gamma V(S_{t+1}) \mid S_t = s) \\ q(s) = E(R_{t+1} + \gamma q(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a) \end{cases}$$

7. Erläutern Sie was der Vorteil von Reinforcement Learning mit Funktionsapproximation z.B: durch neuronale Netze gegenüber tabellarischem Reinforcement Learning ist.
Compared with tabular reinforcement learning, neural network can approximate complex value functions with continuous or high dimensional states and actions.
eg. in autonomous driving it's often used.

8. Nennen Sie zwei Gründe, warum die Berechnung der Zustandswertfunktion für die optimale Policy V ∗(s) rekursiv durchgeführt werden sollte, anstatt mit einer direkten Methode.
Decompostion of decision-making, handling infinite horizons.

9. On-policy and Off-policy

On-policy: evaluate or improve the policy that is used for experience collection

Vorteil: better exploration Nachteil: high variance

Off-policy: evaluate or improve a target policy which is different from behavior policy used for experience collection.

Chapter 10 NLP

1.Goal of NLP

Develop models and algorithms that can understand and generate natural language

2.Text normalization (Preprocessing)

Convert text into a simple and standardized form that can be processed and understood by computer.

Method: lowercasing, removing step words

3. Tokenization

Breaks down a continuous stream of text into small units, such as characters, words , phrases or sentences called token. (Capture meaningful foundamental syntactic and semantic element of language)

a).White space tokenization(Word)

Split text based on white space, problem: out of vocabulary word

b).Byte Pair tokenization(Subword)

Deal with unknown words(low+new+newer-lower), frequently appearing words are compressed to one token, rarely appearing words are converted into multiple tokens.

4.Text representation

Encoding the meaning of word to make them computer-understandable

a).One-hot encoding: sparse and high dimensional vectors, no similarity relations between vectors

b).Word2vec(generate word vector):

CBOW: from context words derive central words(hyperparameter half size c)

Skip-gram: from central words derive contex words

CBOW Input: $\frac{1}{2C}(w_{t-C} \cdots w_{t-1} w_{t+1} \cdots w_{t+C})$

Input ($V \times 1$) → W_1 ($N \times V$) → Hidden layer ($N \times 1$) → W_2 ($V \times N$) → output ($V \times 1$)

V: vocabulary size

N: Embending size

word vector: { column of W_1 / row of W_2 / $0.5(W_1 + W_2^T)$

By-product: Word embeddings

5.Language model

Given history context predict the next token.

a). n-gram model(traditional)

Approximate the history context by n-1 tokens

$P(x_t | x_1 \cdots x_{t-1}) \xrightarrow[\text{(memoryless)}]{\text{Markov Assumption}} P(x_t | x_1 \cdots x_{t-(n-1)})$

Problems:

Text is incoherent for small n

Increasing n worsens sparsity problem and increases model size

Huge memory and training set are needed

Limited history context and parameter goes exponentially

b). RNNs(neural network)

Recurrent connection consider not only the current input but also a hidden state that summarize the past inputs for output

Advantages:

Allow processing variable length, same parameter are applied on each step, parameter don't grow with sequence length.

Truncated backpropagation through time: Propagate gradient only back for k time steps instead of all the way to beginning.

Solution of exploding gradient: gradient clipping(clip the gradient, if its norm exceeds the threshold)

Solution of vanishing gradient: Long and short term memory rnns(cell state(stores long term information conceptually like RAM)+hidden state, three gates: forget, input and output)

6.Decoding

Generate human-readable text from machine-readable representation.

Decoding algorithm defines a function g, which select a token from distribution p.

a). Greedy decoding

Select the token with highest probability at each step, token is only locally optimal, but might not be the best choice in long term

b). Exhaustive search

Computation all possible sequences, take the one that maximize the probability, computationally expensive

c).Beam search

On each step, only keep track of k most probable translations, more efficient but no garantee to find the optimal

7. Neural text degeneration(solution of repetition)

Naive sampling, top-k, top-p