Unsupervised Learning

COMP9417, 23T2

Section 1

Unsupervised Learning

Unsupervised Learning

Learning without any labels.

For example,

- Cluster analysis (i.e grouping users of a social media, classifying similar events/data without knowing any other information)
- Signal separation (i.e PCA, SVD)

Unsupervised Learning

Learning without any labels.

For example,

- Cluster analysis (i.e grouping users of a social media, classifying similar events/data without knowing any other information)
- Signal separation (i.e PCA, SVD)

The content this week is light, so I'll go straight to the lab to explain it.

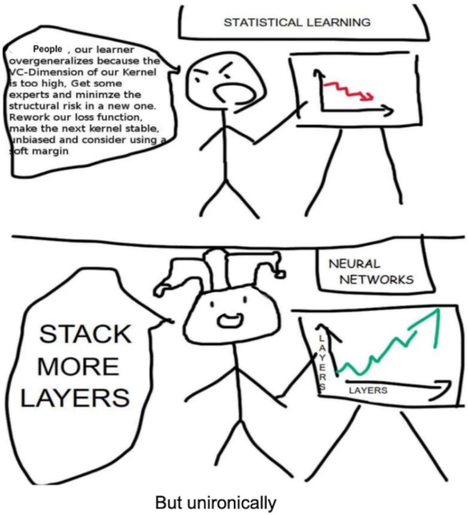
Section 2

The Missing Learning Theory Tut

The Missing Learning Theory Tut

But unironically

The Missing Learning Theory Tut



I'll focus on PAC learning and VC dimension, but we also introduce the WINNOW algorithms and the *No Free Lunch theorem* in lectures.

Section 3

PAC Learning

PAC Learning

Probably **A**pproximately **C**orrect (PAC) learning is a concept which allows us to quantify whether a learning algorithm will achieve low *true* error.

PAC Learning

Probably **A**pproximately **C**orrect (PAC) learning is a concept which allows us to quantify whether a learning algorithm will achieve low *true* error.

We typically see the example of binary classification here and set the problem as follows:

- Observed data D sampled from a true distribution $\mathcal{D}$
- A *concept* c generates the data i.e we have data $D = \{(x_i, c(x_i))\}_{i=1}^{m}$
- We aim to model the concept using a hypothesis h from a class of hypotheses H

PAC Learning

Probably **A**pproximately **C**orrect (PAC) learning is a concept which allows us to quantify whether a learning algorithm will achieve low *true* error.

We typically see the example of binary classification here and set the problem as follows:

- Observed data D sampled from a true distribution $\mathcal{D}$
- A *concept* c generates the data i.e we have data $D = \{(x_i, c(x_i))\}_{i=1}^m$
- We aim to model the concept using a hypothesis h from a class of hypotheses H

In this setting, we define the true error of a hypothesis as,

$$\text{Err}_{\mathcal{D}}(h) = \Pr_{x \in \mathcal{D}}(c(x) \neq h(x))$$

The **version space** of a learning problem is a subspace of H where the training error (denoted r) is 0.

The **version space** of a learning problem is a subspace of H where the training error (denoted r) is 0.

We say a version space is ϵ-exhausted if for all hypotheses in that space, we have less than ϵ true error i.e $(\forall h \in VS_{H,D})\text{Err}_{\mathcal{D}}(h) < \epsilon$.

The **version space** of a learning problem is a subspace of H where the training error (denoted r) is 0.

We say a version space is ϵ-exhausted if for all hypotheses in that space, we have less than ϵ true error i.e $(\forall h \in VS_{H,D})\text{Err}_{\mathcal{D}}(h) < \epsilon$.

How many training examples do we need to ϵ-exhaust the version space for a problem?

Say H is finite, and D is a sequence of m independent random samples of a concept c.

What is the probability of a hypothesis in the version space having an error greater than or equal to ϵ?

Say H is finite, and D is a sequence of m independent random samples of a concept c.

What is the probability of a hypothesis in the version space having an error greater than or equal to ϵ?

If $h \in VS_{H,D}$,

$$\Pr(\mathrm{Err}_{\mathcal{D}}(h) \geq \epsilon)$$

Say H is finite, and D is a sequence of m independent random samples of a concept c.

What is the probability of a hypothesis in the version space having an error greater than or equal to ϵ?

If $h \in VS_{H,D}$,

$$\text{Pr}(\text{Err}_{\mathcal{D}}(h) \geq \epsilon) = (1 - \epsilon)^m$$

Say H is finite, and D is a sequence of m independent random samples of a concept c.

What is the probability of a hypothesis in the version space having an error greater than or equal to ϵ?

If $h \in VS_{H,D}$,

$$\Pr(\mathrm{Err}_{\mathcal{D}}(h) \geq \epsilon) = (1 - \epsilon)^m$$

by definition,

$$(1 - \epsilon)^m < e^{-\epsilon m}$$

Say H is finite, and D is a sequence of m independent random samples of a concept c.

What is the probability of a hypothesis in the version space having an error greater than or equal to ϵ?

If $h \in VS_{H,D}$,

$$\Pr(\mathrm{Err}_{\mathcal{D}}(h) \geq \epsilon) = (1-\epsilon)^m$$

by definition,

$$(1-\epsilon)^m < e^{-\epsilon m}$$

Therefore,

$$\Pr(\mathrm{Err}_{\mathcal{D}}(h) \geq \epsilon) < |H|e^{-\epsilon m} \text{ for all } h \in H$$

We can then bound this probability by some $0 \leq \delta \leq 1$.

$$|H|e^{-\epsilon m} \leq \delta$$

We can then bound this probability by some $0 \leq \delta \leq 1$.

$$|H|e^{-\epsilon m} \leq \delta$$

Using simple log laws, we get

$$m \geq \frac{1}{\epsilon}(\ln(|H|) + \ln(1/\delta))$$

We can then bound this probability by some $0 \leq \delta \leq 1$.

$$|H|e^{-\epsilon m} \leq \delta$$

Using simple log laws, we get

$$m \geq \frac{1}{\epsilon}(\ln(|H|) + \ln(1/\delta))$$

We now have a bound of the number of examples needed to assure that $(\forall h \in VS_{H,D})\text{Pr}(\text{Err}_{\mathcal{D}}(h) \leq \epsilon) \geq 1 - \delta$.

We say a concept class C is PAC-learnable by a learner L using a hypothesis class H for all $c \in C$ and distributions $\mathcal{D}$ if for all $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$ the learner outputs a hypothesis $h \in H$ such that,

$$\mathsf{Err}_{\mathcal{D}}(h) \leq \epsilon$$

with probability $1 - \delta$.

We say a concept class C is PAC-learnable by a learner L using a hypothesis class H for all $c \in C$ and distributions $\mathcal{D}$ if for all $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$ the learner outputs a hypothesis $h \in H$ such that,

$$\text{Err}_{\mathcal{D}}(h) \leq \epsilon$$

with probability $1 - \delta$. In time that is polynomial in $1/\epsilon$, $1/\delta$, m and $|C|$.

Section 4

VC Dimension

VC Dimension

How do we measure model ‘complexity’?

VC Dimension

How do we measure model ‘complexity’? Vapnik and Chervonenkis had the same question.

First, we define a **dichotomy** of a set as a partitioning of that set into two disjoint subsets.

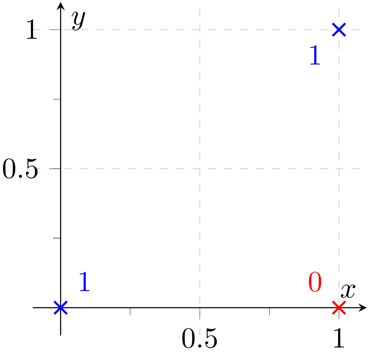
VC Dimension

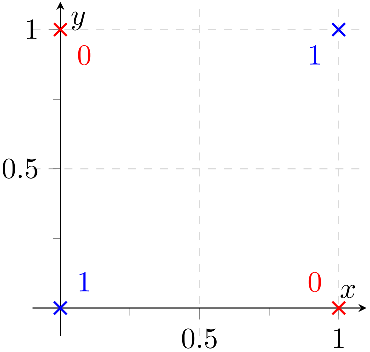
How do we measure model ‘complexity’? Vapnik and Chervonenkis had the same question.

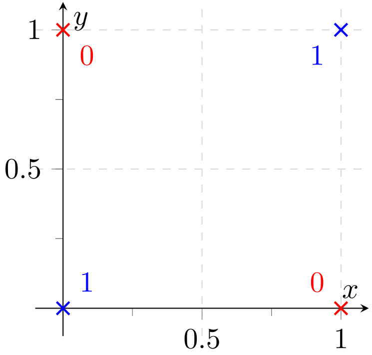
First, we define a **dichotomy** of a set as a partitioning of that set into two disjoint subsets.

We also say a set is **shattered** by a hypothesis space if for every dichotomy there is a hypothesis from that space which is consistent with that dichotomy.

Lots of big words, what does it mean?




y
1
0.5
0
1
1
0
x
0.5
1



We can't shatter this dataset with the space of linear classifiers!

The VC-Dimension of a hypothesis space is the size of the largest finite subset of an instance space $\mathcal{X}$ which can be shattered by that hypothesis space (typically denoted $VC(H)$).

The VC-Dimension of a hypothesis space is the size of the largest finite subset of an instance space $\mathcal{X}$ which can be shattered by that hypothesis space (typically denoted $VC(H)$).

For the previous example we have $VC(H) = 3$, though for more complex hypothesis classes we can have $VC(H) \equiv \infty$.

The VC-Dimension of a hypothesis space is the size of the largest finite subset of an instance space $\mathcal{X}$ which can be shattered by that hypothesis space (typically denoted $VC(H)$).

For the previous example we have $VC(H) = 3$, though for more complex hypothesis classes we can have $VC(H) \equiv \infty$.

We can also generalise the bound of m from PAC-learning to include possibly non-finite hypothesis classes,

$$m \geq \frac{1}{\epsilon}(4\log_2(2/\delta) + 8VC(H)\log_2(13/\epsilon))$$

That's it for the term, good luck in the exam period! 😄

Do myExperience, study hard etc. etc.