بسم الله الرحمن الرحيم .. نبدأ محاضره جديده

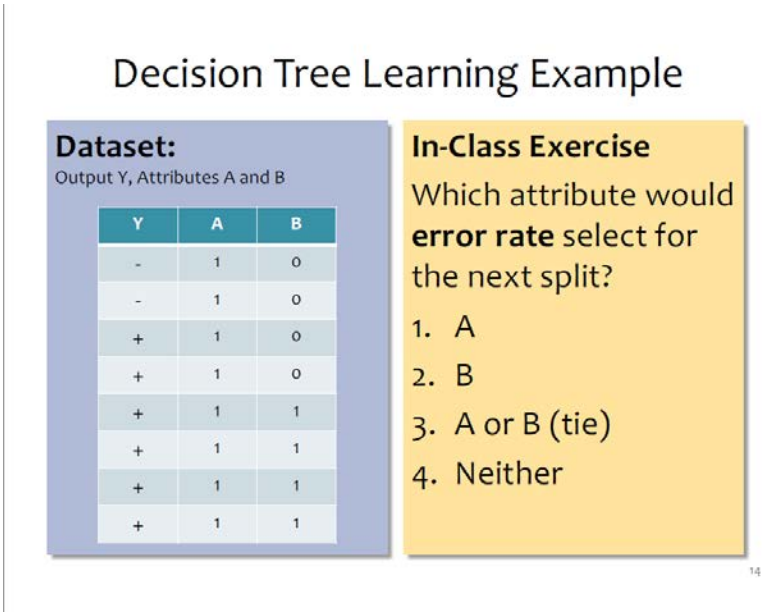
هنتكلم شويه علي high level topics .. هنتكلم علي ال overfitting and K-Nearest neighbors .. الدكتور اتكلم شويه علي ال staff .. ناس جميله والله ربنا يوفقهم و يوفقنا جميعاً .. الدكتور قال ان في حد سأل سؤال في الكلاس ..

When and how do we decide to stop growing trees? What if the set of values an attribute could take was really large or even infinite?

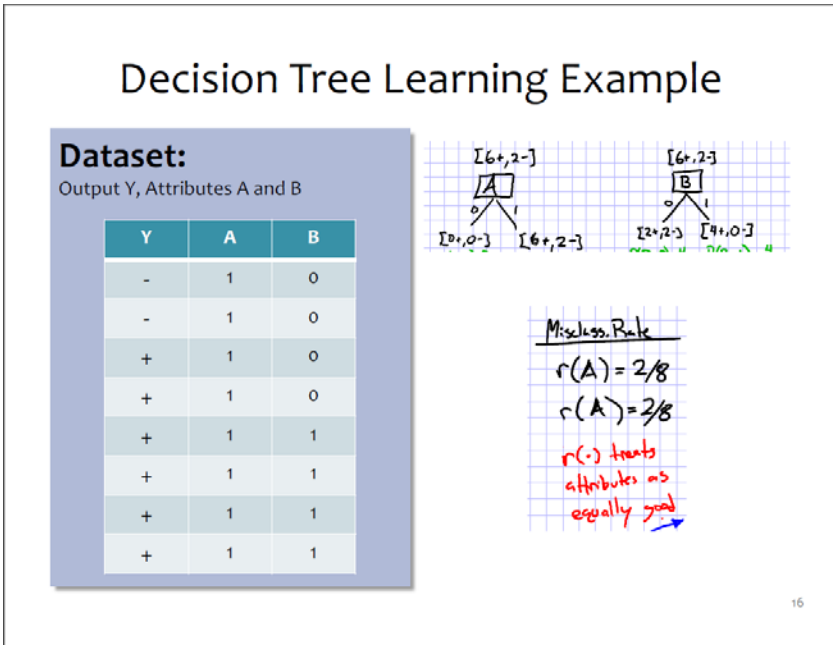
انهارده هنتكلم انهارده يعني علي حتة ال discrete attributes .. بس لو كان في attribute و كان real value .. كمية الارقام اللي ممكن تاخدها هتبقا كتيره جداً جداً .. بس في حركة كدا بت consider ال order بتاع ال split اللي بيحصل في ال tree بيقا O(L) ... و ال L هنا هو عدد ال values الي ال attribute بياخدها في ال training set .. فبيقلك فكّر إيه هي ال trick اللي ممكن تبقا موجوده؟

....

آخر مره احنا اتكلمنا عن ال splitting criteria for decision trees .. وقلنا ان ال splitting criteria هي ال how we measure the effectiveness of a particular split .. في الصوره اللي جايه ديه :: احنا قلنا في المحاضره ان كان في تعادل ما بين ال split on A or B ...



الحل اهو::



وبعدين اتكلمنا علي ال Gini impurity و إنها different splitting criterion .. اللي هي Gini gain .. وبعدين الدكتور كان قال اننا نعملها علي الداتا سيت بتاعتنا .. بس انا معملتش حاجه .. فتعال نشوف الدكتور عمل ايه ..

أول حاجه عند ال Root node .. G(Y) = 1 – P(Getting it right as a plus) – P(Getting it right as a minus) فديه هي قيمة ال impurity

فالسؤال هنا .. how much can we reduce that impurity that exists in the root by splitting on either attribute A or B??

تعال نشوف: لو احنا فصلنا علي A .. هنقول إن P(A=0) = 0 وهو إن صفر من ال Training examples عندها الحاله ديه ... و P(A=1) = 1 اللي هو كلو هيخش في الناحيه ديه .. طيب لو جينا نقول بقي:

G(Y| A=1) = G(Y and A) / A = G(Y)

كدا بقت نفس الرووت نوود .. انما لو قلت الكوندشن عند A=0 .. انت معندكش داتا اصلا لل A=0 .. فالدكتور سأل سؤال .. ازاي نقدر ن

Estimate a quantity about the random variable Y if we have no data about Y?????

فبتسمي الحاله ديه ... undefined

انت لما بتيجي تحسب ال Gini gain بتعمل ايه .. بتقول آدي عندك القيمه اللي هي 0.375 عند الرووت .. و تطرح منها ال weighted average of the gini impurity at the 2 leaf nodes

ده اللي هو undefined * 0 و ال 1 * 0.375 .. و ده طلعلك صفر Gini gain .. انت كدا مقللتش ال impurity نهائي ..

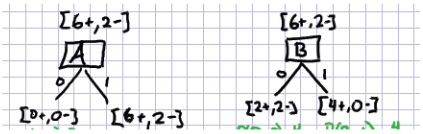
تعال نشوف حالة ال B ... P(B=1) = P(B=0) .. ولما جيت تتحرك زي A .. لاقيت ان عندك reduction in impurity = 0.125 لما جيت تفصل عند B .. فعند ال root ال impurity كانت 0.375 .. بس ال weighted average ما بين ال 2 leafs .. بقي 0.25.. فال Gini gain هنا عمل ايه .. هو خد some probabilistic notion لل remaining decisions اللي انت عاوز تعملها .. فانت بتقول ايه .. بتقول انك عاوز attribute بتقلل ال impurity on both sides ... وفي فرق ما بين انك تفصل جزئياً ال impurity و انك تفصلها كلياً ... و الدكتور بيقول اننا بن capture الحته ديه باستخدام ال expected miss classification rate ... هو قال كمان ان ال error rate احنا بس بنبص علي pure error rate .. انما هنا بنتكلم علي probabilistic notion of error rate ... فهنا في المثال ده .. في attribute فاز اللي هو B ...

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1



1) G(Y) = 1 – (6/8)² – (2/8)² = 0.375

2) P(A=1) = 8/8 = 1
3) P(A=0) = 0/8 = 0
4) G(Y | A=1) = G(Y)
5) G(Y | A=0) = undef
6) GiniGain(Y | A) =
0.375 – 0(undef) – 1(0.375) = 0

7) P(B=1) = 4/8 = 0.5
8) P(B=0) = 4/8 = 0.5
9) G(Y | B=1) = 1 – (4/4)² – (0/4)² = 0
10) G(Y | B=0) = 1 – (2/4)² – (2/4)² = 0.5
11) GiniGain(Y | B) =
0.375 – 0.5(0) – 0.5(0.5) = 0.125

20

الدكتور دلوقت هيتكلم علي mutual information ... فكر في 2 راندم فاريبلز إكس و واي .. المهم لما جيه يحسب ال entropy .. حسبه كإنه expected log probability of a random variable .. الدكتور بيقول إن ال entropy هو شبه اللي كنا بنعملو في ال Gini impurity .. هو بيقيس the sort of stochasticity in a collection of data .. فاللي هنحاول نعملو اننا نقلل ال Entropy عن طريق اننا نختار attribute معين .. أول حاجه ممكن نعملها هو إننا نحدد a specific conditional entropy .. وديه شبه ال Gini impurity on one of the leaf nodes..الدكتور بيقول اننا هنستخدم ال Specific conditional entropy كإنها entropy بعد ما انت خلاص عملت ال condition علي ال attribute انو ياخد قيمه معينه .....

ال Conditional entropy ده بقا هو بياخد weighted average of specific conditional entropy .. بمعني تاني ديه احتمالية إن ال attribute ياخد قيمه x مضروبه في ال specific conditional entropy for that x .. اللي هو فكّر فيها كإن لكل leaf node احنا عملنا weight علي حسب عندنا كام مثال مشي في ال branch ده .. وبعدين خد ال specific conditional entropy .... بعد كدا ال mutual information .. بيتحسب بال entropy at root node تطرح منو ال weighted average of entropies after conditioning on X ... والدكتور قال ان ده اللي كنا بنعملو في ال Gini Gain .. احنا هناخد ال Impurity at root – weighted average of the imuprities .. فلو بنتكلم علي ال decision tree احنا نقدر نستخدم ال definition ده بتاع ال mutual information ك splitting criterion

Mutual Information

Let X be a random variable with $X \in \mathcal{X}$. ← attribute
Let Y be a random variable with $Y \in \mathcal{Y}$. ← class ← expected log probability

Entropy: $H(Y) = -\sum_{y \in \mathcal{Y}} P(Y = y) \log_2 P(Y = y)$

Specific Conditional Entropy: $H(Y \mid X = x) = -\sum_{y \in \mathcal{Y}} P(Y = y \mid X = x) \log_2 P(Y = y \mid X = x)$

Conditional Entropy: $H(Y \mid X) = \sum_{x \in \mathcal{X}} P(X = x) H(Y \mid X = x)$

Mutual Information: $I(Y; X) = H(Y) - H(Y|X) = H(X) - H(X|Y)$

- For a decision tree, we can use **mutual information** of the output class Y and some attribute X on which to split **as a splitting criterion**
- Given a dataset D of training examples, we can estimate the required probabilities as...

$P(Y = y) = N_{Y=y}/N$
$P(X = x) = N_{X=x}/N$
$P(Y = y|X = x) = N_{Y=y,X=x}/N_{X=x}$

where $N_{Y=y}$ is the number of examples for which $Y = y$ and so on.

22

الانتروبي هو high level بنقيس بيه ال expected #of bits عشان نقدر ن code 1 random draw from X فكّر فيها كإنها قياس ل level of uncertainty encoded in a random variable..فمثلاً لو كان random variable كان بيتكلم علي coin flip و كان دايماً بيجي ال head .. فهنا مفيش random uncertainty .. يعني لو جيت ت flip تاني .. هيطلع head .. انما لو flipped the coin multiple times .. ونص الوقت بيطلع هيد و النص التاني بيطلع tails .. فهنا انت عندك high uncertainty .. فبالتالي high entropy لإنك مش عارف ايه اللي جاي المره الجايه .. فبالتالي احنا في ال decision tree .. عاوزين نقلل ال entropy of the random variable that we try to predict ..

Mutual Information

Let X be a random variable with $X \in \mathcal{X}$.
Let Y be a random variable with $Y \in \mathcal{Y}$.

Entropy: $H(Y) = -\sum_{y \in \mathcal{Y}} P(Y = y) \log_2 P(Y = y)$

Specific Conditional Entropy: $H(Y \mid X = x) = -\sum_{y \in \mathcal{Y}} P(Y = y \mid X = x) \log_2 P(Y = y \mid X = x)$

Conditional Entropy: $H(Y \mid X) = \sum_{x \in \mathcal{X}} P(X = x) H(Y \mid X = x)$

Mutual Information: $I(Y; X) = H(Y) - H(Y|X) = H(X) - H(X|Y)$

- **Entropy** measures the **expected # of bits** to code one random draw from X.
- For a decision tree, we want to **reduce the entropy of the random variable we are trying to predict!**

Conditional entropy is the expected value of specific conditional entropy
$E_{P(X=x)}[H(Y \mid X = x)]$

Informally, we say that **mutual information** is a measure of the following:
If we know X, how much does this reduce our uncertainty about Y?

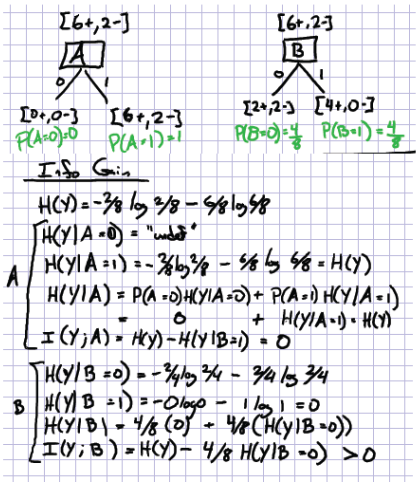
لو رجعنا لنفس الداتا سيت تاني .. تقدر تحسب ال mutual information as a splitting criterion ..

Decision Tree Learning Example

Dataset:

Output Y, Attributes A and B

Y	A	B
-	1	0
-	1	0
+	1	0
+	1	0
+	1	1
+	1	1
+	1	1
+	1	1



26

الدكتور دلوقت عاوز يتكلم علي Comparison of splitting criteria .. في ال practice .. انهي واحده من الطرق ديه احسن حاجه .. ازاي نقدر ن work through which of these works best on datasets

الدكتور بيستعرض داتا سيت من ال 1992 .. كانت بتتكلم علي 4 splitting criterion on 12 datasets ....

Experiments: Splitting Criteria

Bluntine & Niblett (1992) compared 4 criteria (random, Gini, mutual information, Marshall) on 12 datasets

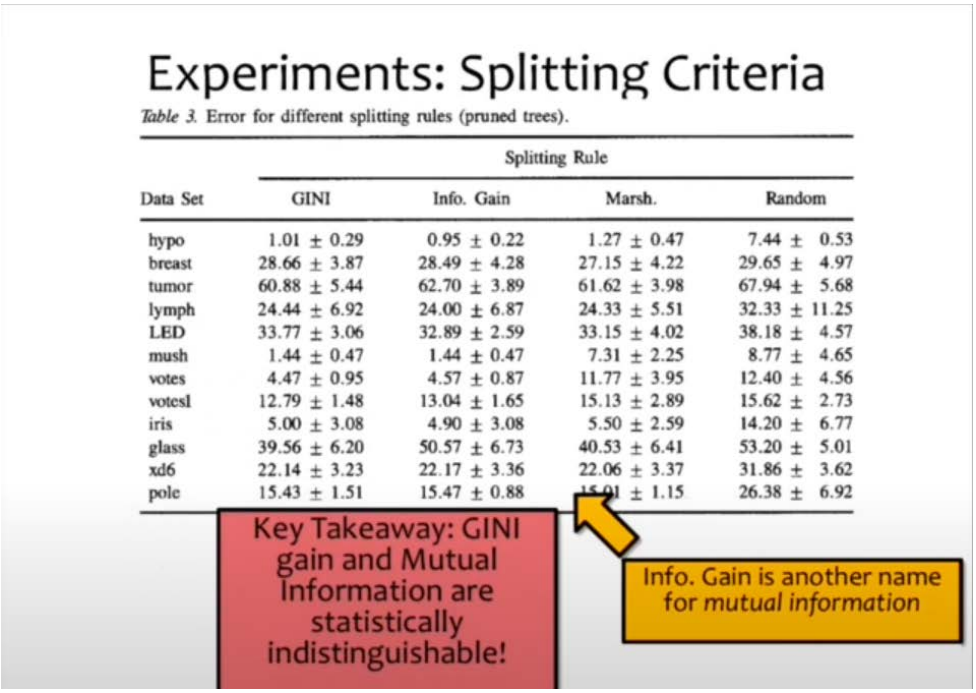
Medical Diagnosis Datasets: (4 of 12)

- **hypo**: data set of 3772 examples records expert opinion on possible hypo- thyroid conditions from 29 real and discrete attributes of the patient such as sex, age, taking of relevant drugs, and hormone readings taken from drug samples.
- **breast:** The classes are reoccurrence or non-reoccurrence of breast cancer sometime after an operation. There are nine attributes giving details about the original cancer nodes, position on the breast, and age, with multi-valued discrete and real values.
- **tumor:** examples of the location of a primary tumor
- **lymph**: from the lymphography domain in oncology. The classes are normal, metastases, malignant, and fibrosis, and there are nineteen attributes giving details about the lymphatics and lymph nodes

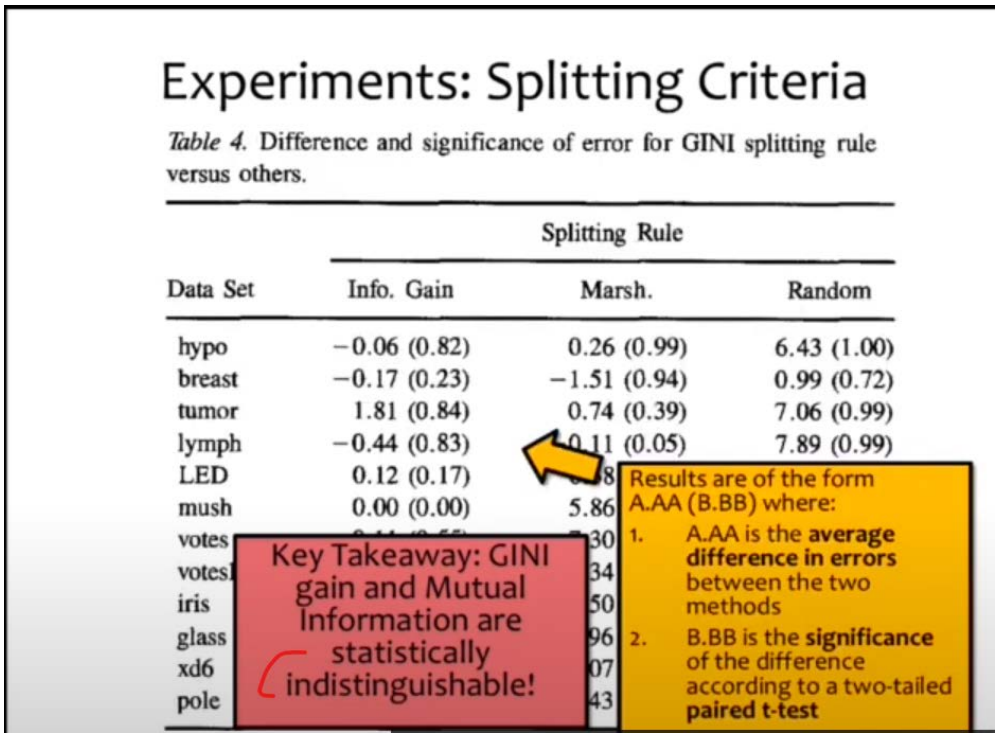
Table 1. Properties of the data sets

Data Set	Classes	Attr.s	Training Set	Test Set
hypo	4	29	1000	2772
breast	2	9	200	86
tumor	22	18	237	102
lymph	4	18	103	45
LED	10	7	200	1800
mush	2	22	200	7924
votes	2	17	200	235
votes1	2	16	200	235
iris	3	4	100	50
glass	7	9	100	114
xd6	2	10	200	400
pole	2	4	200	1647

ديه بتقارن ما بين 4 splitting criterion علي 12 داتا سيت .. ال 12 دااتا سيتس مكتوبين عندك اهم .. احنا هنتكلم علي أول 4 .. عشان هم medical diagnosis datasets .. احنا هنركز علي Gini gain and Mutual information ....



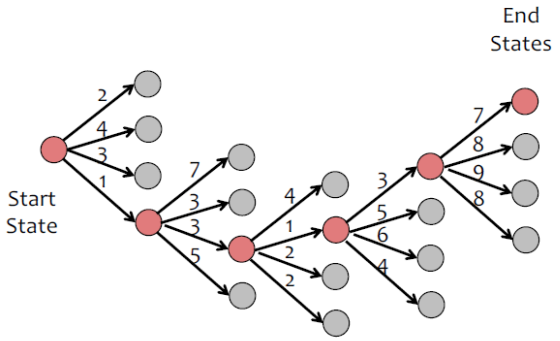
ديه الإيرورز .. والصوره الجايه ديه بتعبر عن تفاصيل أكتر من ناحية ال statictical significance of the error difference between Ginny gain and mutual information



أول رقم هو ال average difference و تاني رقم هو ال significance of that differenceده اللي هو يعني P – 1 .. أي حاجه أعلي من 0.95 هتبقا statistically significant difference .. لما هتيجي تبص علي أول عمود اللي هو Info. Gain .. هتلاقي ان مفيش ولا واحده من الداتا سيتس significatlly difference .. فاللي ناخدو من الحوار ده ان مفيش فرق ما بين الاتنين لما جم يقارنو ما بينهم ..

دلوقت هنعوز نتكلم علي inductive bias لل decision trees .. تعال دلوقت نتكلم علي ال algorithmic prespective الي ممكن تبقا مختلفه عن ايه اللي ممكن نعمله .. ايه هو ال greedy search .. مراجعه سريعه عليه .. عندك سيرش سبيس .. بتتكون من نودز و edges وال edges ديه معاها weights .. و الهدف ان نلاقي lowest total weight path from a root node to a leaf node ... فال greedy search بت follow البروسيس ديه .. عند كل نوود اختار ال edge اللي ليها lowest immediate weight .. خد النود ديه و بعدين عيد تاني .. .. لو فضلت ماشي كدا هتوصل فعلاً للآخر بس بص علي اللي هيحصل في الصور الجايه .. وده حصل عشان انت طماع .. دايماً عاوز أقل حاجه

Background: Greedy Search

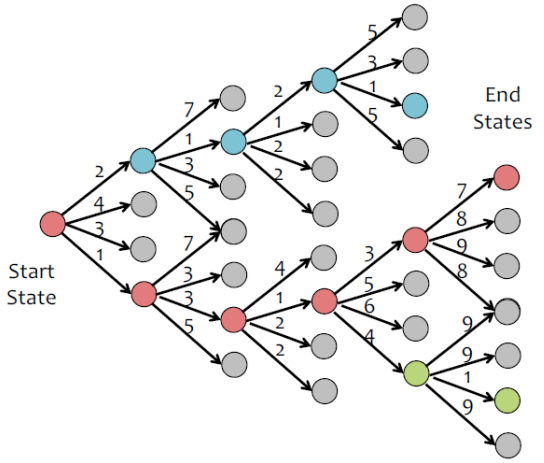


Goal:
- Search space consists of nodes and weighted edges
- Goal is to find the lowest (total) weight path from root to a leaf

Greedy Search:
- At each node, selects the edge with lowest (immediate) weight
- Heuristic method of search (i.e. does *not* necessarily find the best path)

38

Background: Greedy Search



Goal:
- Search space consists of nodes and weighted edges
- Goal is to find the lowest (total) weight path from root to a leaf

Greedy Search:
- At each node, selects the edge with lowest (immediate) weight
- Heuristic method of search (i.e. does *not* necessarily find the best path)

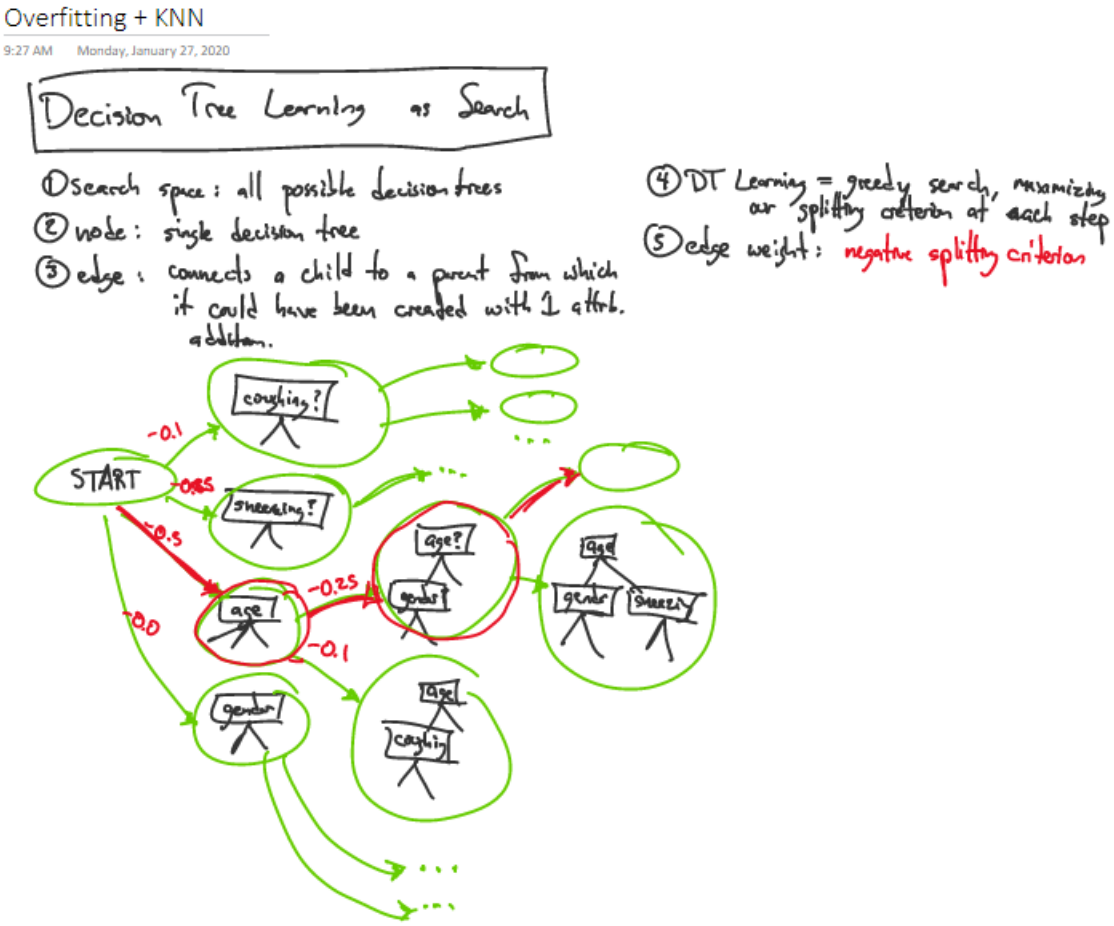
40

فال جريدي سيرش .. هي مش شرط تلاقي أحسن طريق .. عشان هي هترجعلك الطريق الاحمر .. بس في طرق تانيه انت مدخلتهاش عشان انت علي طول طماع عاوز دايماً تكسب في اللحظه ديه .. انت عادي ممكن تخسر النوود ديه بس هتكسب مكانها قدام ... وده اللي بيحصل ف الدنيا عموماً ساعات بتقول هي ديه .. هي ديه .. بس مبتبقاش هي ... فأسلم حل .. سلّمها لله و توكل علي الله .. شفت بقي لما هتفهم وتلاقي الكونسبت شغال هتقتنع ازاي .. توكّل علي الله وهو اللي هيفتحلك الطريق اللي انت مش شايفو ..

دلوقت احنا عاوزين نفكّر في ال decision tree learning as search .. عاوز نفهم ازاي نقدر ناخد المشكله اللي شغالين عليها في ال decision tree و ن recast as a search problem .. أول حاجه هنفكر فيها هو ايه هو ال entire search space بتاعنا ..

1. السيرش سبيس بتاعنا هو كل الاحتمالات المتاحه لكل ال decision trees
2. Node: كل نوود عباره عن decision tree
3. Edge: كل إيدج هتوصّل a child ل ال parent from which it could have been created with 1 additional attribute

تعال نبص ع الصوره ديه . الارقام بالسالب خد اقل رقم اللي هو lowest weight path عشان ت maximize the spletting at each step



الصوره ديه هي هي نفس ال decision tree algorithm اللي اتشرح من أول محاضره ... هو هو بس الفرق بتبصلو ازاي .. احنا هنا بنبصلو ك greedy search maximizing our splitting criterion ... واحده من المميزات هنا في طريقة التفكير ديه ان في سؤال مهم جداً ... سؤال جذري في الماشين ..

How is it that your ML algorithm can generalize to unseen examples?

ده السؤال المهم جداً .. خلينا نأجل الرد عليه ثواني ...

تعال نتكلم دقايق بس :

هنحط ديفينيشن بس ... السؤال بتاعنا مرتبط بيه يعني ... ال inductive bias هو ال principal اللي من خلاله بيحصل generalization to unseen examples .. فلما الدكتور سأل انهي tree احنا هنلاقيها في الألجورذم بتاعنا .. هو كان يقصد ايه هو ال inductive bias بتاع ال decision tree learning .. ايه هو ال principle اللي هو بيشتغل بيه عشان يجيب شجره كويسه .. من ضمن الشجر الكتير اللي فوق .. ازاي اختار ما بينهم ... حد رد وقالو اختار أقل عمق للشجره ... نقدر نقول ان ال inductive bias هيحاول يختار اصغر شجره ت match the data وليها أعلي mutual information attributes ...

DT: Remarks

ID3 = Decision Tree Learning with Mutual Information as the splitting criterion

Question: Which tree does ID3 find?

Definition:
We say that the **inductive bias** of a machine learning algorithm is the principal by which it generalizes to unseen examples

Inductive Bias of ID3:
Smallest tree that matches the data with high mutual information attributes near the top

Occam's Razor: (restated for ML)
Prefer the simplest hypothesis that explains the data

44

دلوقت الدكتور بيقولك عاوزنا نفكّر في ألجورذم مختلف شويه عن اللي احنا فيه .. احنا هنا هنعمل exhaustive global search .. بمعني اننا هنعوز ن enumerate كل شجره موجوده في السيرش سبيس .. كل اتحمال موجود فكر فيه .. وبعدين انسي خالص ال edge weights ... وبعدين إسأل نفسك لكل شجره انهي شجره ليها أقل training error و ديه هتبقا احسن حاجه .. بس ممكن يكون في اكتر من شجره بيدوك 0 ايرور .. ازاي اختار احسن واحده منهم؟