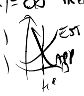
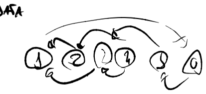
1. tWhat is the inductive bias in machine learning? REDUCING THE SEARCH SPACE FOR ERM
2. What does it mean that machine learning is distribution-free? IT works $\forall D$ over X
3. Definition of PAC-learning H PAC IF UNDER $f \in H$ AND 0-1 LOSS $\exists A, \exists m_H \mid \forall D$ over $X, \forall f \in H, \forall \varepsilon, \delta \in (0,1)$ if S iid with $|S| \geq m_H(\varepsilon, \delta)$ then $L_D(S) \leq \varepsilon$ wp $1-\delta$
4. What is the ERM? $ERM_H \in \arg\min_{h \in H} L_S(h)$
5. What is the realizability assumption? $f \in H$
6. What is the sample complexity? MINIMUM NUMBER OF IID SAMPLES of X wrt D for which $A(S)$ CAN EFFICIENTLY LEARN
7. Are finite hypothesis classes learnable? What's their sample complexity? yes, $\frac{\log(|H|/\delta)}{\varepsilon}$
8. Differences between Agnostic PAC learnability and PAC learnability in PAC $f \in H$ so $\min_{h \in H} L_D(h) = 0$ because $h^* \in H \mid h^* = f$
9. Definition of Agnostic PAC learnability
10. What is in general the expected loss of a predictor? $L_D(h) = \mathbb{E}[\ell(h,(x,y))]$
11. What is uniform convergence? UC: if $\exists m_H^{UC}: (0,1)^2 \to \mathbb{N} \mid$ wrt loss ℓ, $\forall \varepsilon, \delta \in (0,1)$ if S iid from X with D then if $|S| \geq m_H^{UC}(\varepsilon, \delta)$ S is ε representative wp $1-\delta$
12. If a hypothesis class has the uniform convergence property, what can you tell me about its learnability? UC $\to S$ ε representative $\to H$ APAC $\Rightarrow H$ PAC
13. What is the sample complexity of a finite hypothesis class? $\frac{\log(|H|/\delta)}{\varepsilon}$
14. What is the discretization trick?
15. Are infinite classes PAC-learnable? And Agnostic PAC-learnable?
16. What are the consequences of the No Free Lunch theorem? if $|X| = \infty$ then $H_{\{0,1\}^X}$ not PAC wrt 0-1 loss
17. What is the bias-complexity trade-off? given $H' \subseteq H$ $\varepsilon_{APP}(H) \leq \varepsilon_{APP}(H')$ $\varepsilon_{EST}(H') \leq \varepsilon_{EST}(H)$



18. What is the approximation error?
19. What is the estimation error? $L_D(A(S)) = \underbrace{L_D(A(S)) - \min_{h \in H} L_D(h)}_{\varepsilon_{EST}} + \underbrace{\min_{h \in H} L_D(h)}_{\varepsilon_{APP}}$
20. Advantages/disadvantages of selecting a more complex hypothesis class ε_{APP} DECREASES BUT ε_{EST} INCREASE. ALSO ERM_H BECOME COMPUTATIONALLY MORE DIFFICULT.
21. What is the Bayes error? $L_D(A(S)) = L_D(A(S)) - \min_{h \in H} L_D(h) + \min_{h \in H} L_D(h) - \min_h L_D(h) + \min_h L_D(h)$ IRREDUCIBLE ERROR GIVEN BY THE DISTRIBUTION OF THE DATA
22. Definition of shattering H shatters $A \subset X$ if $\exists h \in H \mid h(x) = g(x)\ \forall x \in A, \forall g: A \to \{0,1\}$
23. VC dimension $VC(H) = \max |X| \mid X$ shattered by H
24. VC dimension and size of the class if $H \subset H'$ then $VC(H) \leq VC(H')$
25. What if VC-dimension is infinite? $VC(H) = \infty \to H$ not PAC
26. ERM is a successful PAC-learner for a class H, what consequences?



27. ERM is a successful agnostic PAC-learner for a class H, what consequences?
28. What is the sample complexity of a PAC learnable class? $\approx \frac{VC(H) + \ln(1/\delta)}{\varepsilon}$ as of FTSL
29. What is the sample complexity of an agnostic PAC learnable class? $\approx \frac{\quad}{\varepsilon^2}$
30. How many samples are needed for a class to have the Uniform Convergence property? $m \geq m_H^{UC}(\varepsilon, \delta) \approx \frac{d + \ln(1/\delta)}{\varepsilon^2}$
31. Good points and negative points of ERM finds the best predictors but for H too complex can be computationally hard
32. What is the problem with ERM? can be NP-HARD
33. Define a linear classifier $h_{w,b} = \langle w, x \rangle + b = \sum_{i=1}^{m} w_i \cdot x_i + b$
34. Are linear classifiers (PAC/Agnostic PAC) learnable? Why? How? yes in the separable case VC DIM HS is d, in nonh. is $d+1$, we can find the set w of parameter by LP or Perceptron in the separable case.
35. Define linear regression
36. Computational complexity of solving a linear classification problem with the 0-1 loss function ERM_{HS} takes $\Omega(d + m d \log m)$ UNDER
37. Definition of weak learnability H γ-WL if $\exists A: (X,Y)^m \to H$, $\exists m_H^{WL}: (0,1) \to \mathbb{N} \mid \forall D, \forall f \in H, \forall \delta \in (0,1)$ if S iid $|S| = m \geq m_H^{WL}(\delta)$ then $L_D(A(S)) \leq 1/2 - \gamma$ wp. $1-\delta$ with $f \in H$
38. Are there classes that are weak learnable but not PAC learnable? PAC-L $\Leftrightarrow$ WL
39. Why have we discussed weak learnability? By Boosting algorithms we can mixup weak h for B to find h strong for H
40. Describe the AdaBoost algorithm updates $D^{(t)}$, returns $\text{sign}(\sum_{t=1}^{T} w_t h_t)$ FOR $t = 1 \ldots T$ $h_t = WL(S, D^{(t)})$, $\varepsilon_t = L_D(h_t)$, $w_t = \frac{1}{2}\ln(\frac{1}{\varepsilon_t} - 1)$, $D^{(t+1)} \to$ DATA TAKE NOT GUESSED SAMPLES
41. Guarantees for the AdaBoost algorithm
42. What is the effect of the number of iterations in the AdaBoost algorithm?

If WL returns a γ-WLearner h_t then $L_S(h_{ADA}) = e^{-2\gamma^2 T}$ wp $(1-\delta T)$

43. When is a set convex?
44. What is a convex function?
45. How can you check that a function is convex?
46. Why do we care of convexity in ML?
47. When is a ML problem convex? if the loss ℓ is convex w.r.t. h_w $\ell(h_w(x,y))$
48. Linear regression
49. What are surrogate loss functions? convex function that approx. non convex ones
50. Logistic regression
51. What is the optimization error?
52. Definition of non-uniform learnability
53. When is a class non-uniform learnable?
54. What is Structural Risk minimization?
55. What is the minimum description length?
56. What approaches do you know to select the model/the class of hypothesis class?
57. How can we evaluate the quality of a trained model?
58. What is the validation approach to model selection?
59. What is the train-validation-test split?
60. What is k-fold cross validation?
61. What is the model selection curve?
62. What can we do if we are not happy with the quality of the model we trained?

$L_D^{0-1}(A(S)) - L_D^{surr}(A(S)) + L_D^{surr}(A(S)) - \min_{h\in\mathcal{H}} L_D^{surr}(h) + \min_{h\in\mathcal{H}} L_D^{surr}(h) - \min_{h\in\mathcal{H}} L_D^{0-1}(h) + \min_{h\in\mathcal{H}} L_D^{0-1}(h)$

$\mathcal{L}_D$ | ε_{EST} | ε_{OPT} | ε_{APP}

52) H NUC iff $\exists A: (X\times Y)^m \to \mathcal{H}$ $\exists m_{\mathcal{H}}^{UC}: (0,1)\times\mathcal{H}\to\mathbb{N}$, $\forall\delta$, $\forall\varepsilon,\delta\in(0,1)$, $\forall h\in\mathcal{H}$, if S is iid from D of $|S|=m\geq m_{\mathcal{H}}^{UC}(\varepsilon,\delta)$ SAMPLES the $L_D(A(S)) \leq L_S(h)+\varepsilon$ w.p. $1-\delta$

53) H is NUC $\Leftrightarrow$ $\mathcal{H}=\bigcup_n \mathcal{H}_n$ | $VCdim(\mathcal{H}_n)<+\infty$

54) SRM is a non uniform learning algorithm:

Structural Risk Minimization (SRM)

prior knowledge:
$\mathcal{H}=\bigcup_n \mathcal{H}_n$ where $\mathcal{H}_n$ has uniform convergence with $m_{\mathcal{H}_n}^{UC}$
$w:\mathbb{N}\to[0,1]$ where $\sum_n w(n)\leq 1$
define: ϵ_n as in Equation (7.1) ; $n(h)$ as in Equation (7.4)
input: training set $S\sim\mathcal{D}^m$, confidence δ
output: $h\in\operatorname{argmin}_{h\in\mathcal{H}}\left[L_S(h)+\epsilon_{n(h)}(m,w(n(h))\cdot\delta)\right]$

Let $\mathcal{H}=\bigcup_n \mathcal{H}_n$ | $\mathcal{H}_n$ has UC property
Let $w:\mathbb{N}\to[0,1]$ | $\sum_n w_n\leq 1$
Let $\varepsilon_n(m,\delta)=\min\{\varepsilon\in[0,1] \mid m_{\mathcal{H}_n}^{UC}(\varepsilon,\delta)\leq m\} \Rightarrow \min\{\varepsilon\in[0,1]: \varepsilon\geq\sqrt{\frac{VCdim(\mathcal{H}_n)-\ln(1/\delta)}{m}}\}$
Let $n(h)=\min\{n(h)\mid h\in\mathcal{H}_n\}$
output $h\in\operatorname{argmin}_{h\in\mathcal{H}} L_S(h)+\varepsilon_n(m,w_n\cdot\delta) \Rightarrow h\in\operatorname{argmin}_{h\in\mathcal{H}} L_S(h)+\sqrt{\frac{VCdim(\mathcal{H})-\ln(1/w_n\delta)}{m}}$

56) 2 ways:

1) STRUCTURAL RISK MINIMIZATION:
Let $\mathcal{H}=\bigcup_n \mathcal{H}_n$ | $\mathcal{H}_n$ has UC
$w:\mathbb{N}\to[0,1]$ | $\sum_n w_n=1$
$\varepsilon_n(m,\delta)=\min\{\varepsilon\mid m_{\mathcal{H}_n}^{UC}(\varepsilon,\delta)\leq m\}=\min\{\varepsilon\mid\varepsilon\geq\sqrt{c\frac{VCdim(\mathcal{H}_n)+\ln(\frac{1}{\delta})}{m}}\}$
$n(h)=\min\{n\mid h\in\mathcal{H}_n\}$
OUT: $h\in\operatorname{argmin}_{h\in\mathcal{H}} L_S(h)+\varepsilon(m,w_{n(h)}\delta)=\operatorname{argmin}_{h\in\mathcal{H}} L_S(h)+\sqrt{c\frac{VCdim(\mathcal{H}_{n(h)})+\ln(\frac{1}{w_{n(h)}\delta})}{m}}$

2) VALIDATION SET To choose among $\mathcal{H}_1,\dots,\mathcal{H}_m$ we compute $ERM_{\mathcal{H}_i}$ $\forall i=1,\dots,m$
Then we must compare $L_D(h_1),\dots,L(h_m)$ to choose
Not having D we use the VALIDATION SET choosing $\operatorname{argmin}_{i\in\{1,\dots,m\}} L_{VAL}(h_i)$ where $L_{VAL}(h)=\frac{1}{m_{VAL}}\sum_{(x,y)\in VAL}\ell(h(x),y)$
we finally check $|L_{VAL}(h_i)-L_D(h_i)|\leq\sqrt{\frac{\ln(2|\mathcal{H}|)}{2m_{VAL}}}$

59) TS AND TESTS

- given $h = \underset{h'}{\operatorname{argmin}}\ L_{TS}(h')$
- check

$$L_D(h) \approx L_{test}(h) = \frac{1}{m_{tests}} \sum_{(x,y)\in tests} \ell(h,(x,y))$$

- we shall have

$$|L_{tests}(h) - L_D(h)| \le \varepsilon_{test} = \sqrt{\frac{\ln(2|H|/\delta)}{m_{test}}}$$

60) When data is scarce how we apply validation without waste of data?

for $i = 1 - n$ $\quad \{M_1 - M_n\}$

  for $j = 1 - k$ $\quad \{S_1 - S_k\}$

    $h_{ij} = ERM_{H_i}(S_j)$

    $\ell_{ij} = L_{S_j}(h_i)$

  $\ell_i = \frac{1}{k} \sum_{j=1}^{k} \ell_{ij}$

$i^* = \underset{i=1-3}{\operatorname{argmin}}\ \ell_i$

$h = ERM_{H_{i^*}}(S)$