Model selection

$$L_{train} = \frac{1}{n}\sum_{i=1}^{n} L\left(f(x_i^{train}),\ y_i^{train}\right) \leftarrow \text{Empirical Risk}$$

but we want to measure the error on a holdout set.

$$L_{test} = \frac{1}{n}\sum_{i=1}^{n} L\left(f(x_i^{test}),\ y^{test}\right) \leftarrow \text{Expected Risk.}$$

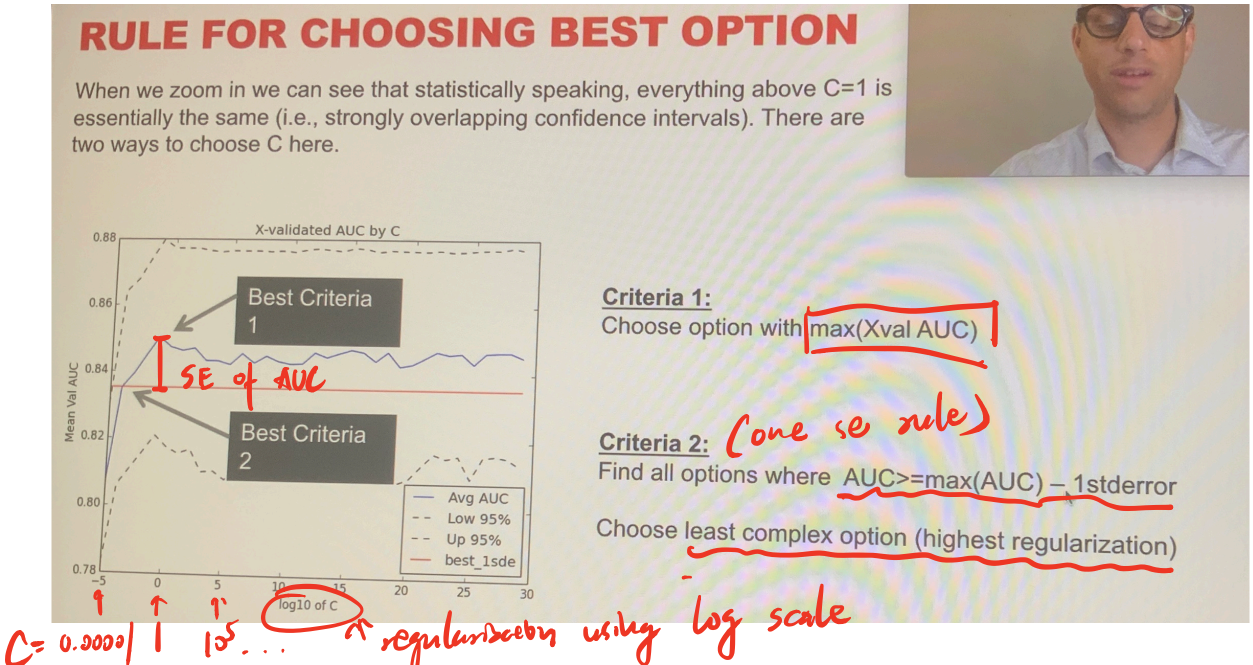
Expected Risk != Empirical Risk

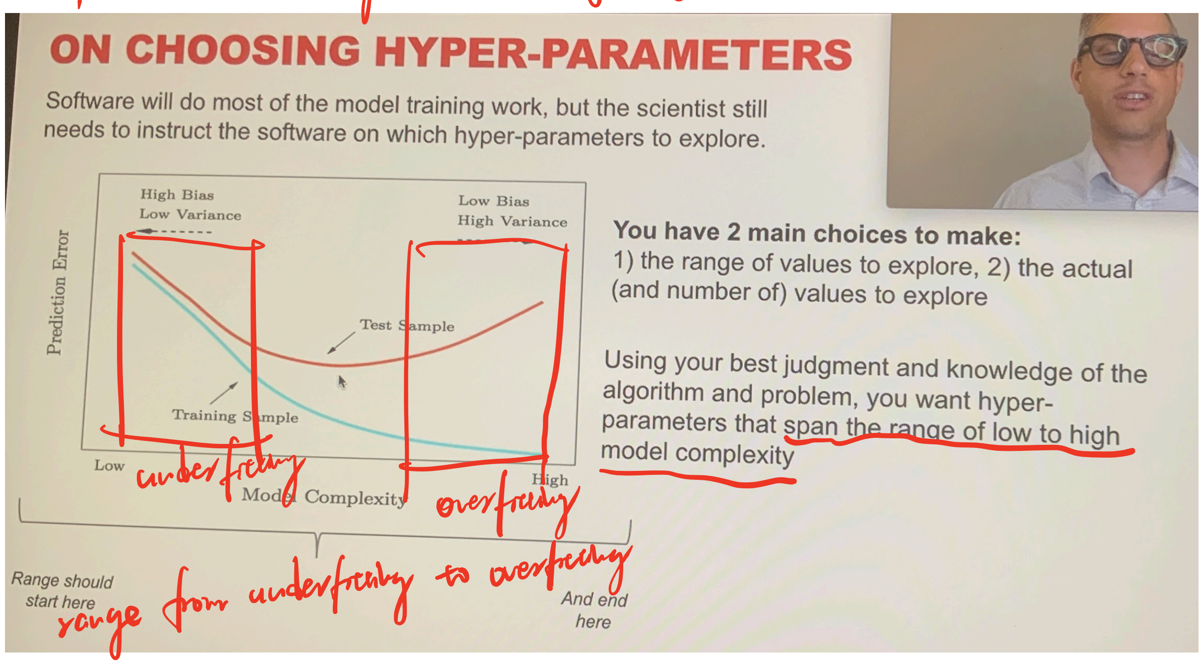
Training Error is our empirical risk and test set error is our best approximation of expected risk.

Example Routine:

Train Data. Validation Data

1. For each configuration c in the set C = [Algorithm × Feature Set × Hyper-parameter]
   - find the function $\hat{f}_c$ such that: $\hat{f}_c = \underset{f \in F}{\operatorname{argmin}} R_{train}$
   - With $\hat{f}_c$ estimated, get the validation loss: $R_c^{val} = \frac{1}{n}\sum_{i=1}^{n} L\left(\hat{f}_c(x_i^{val}), y_i^{val}\right)$
2. Choose the optimal configuration C_{opt} such that: $C_{opt} = \underset{c \in C}{\operatorname{argmax}} R_c^{val}$
3. Define New train = Train + Validation data
4. Find the function $\hat{f}$ such that $\hat{f} = \underset{f \in F}{\operatorname{argmin}} R^{new\ train}$
5. Estimate the test loss as: $R^{test} = \frac{1}{n}\sum_{i=1}^{n} L\left(\hat{f}(x_i^{test}), y_i^{test}\right)$


RULE FOR CHOOSING BEST OPTION
When we zoom in we can see that statistically speaking, everything above C=1 is essentially the same (i.e., strongly overlapping confidence intervals). There are two ways to choose C here.
X-validated AUC by C
Mean Val AUC
0.88
0.86
0.84
0.82
0.80
0.78
-5
0
5
10
15
20
25
30
log10 of C
Best Criteria 1
Best Criteria 2
Avg AUC
Low 95%
Up 95%
best_1sde
SE of AUC
Criteria 1:
Choose option with max(Xval AUC)
Criteria 2: (one se rule)
Find all options where AUC>=max(AUC) – 1stderror
Choose least complex option (highest regularization)
C= 0.0001 1 10^5 ...
regularization using log scale


ON CHOOSING HYPER-PARAMETERS
Software will do most of the model training work, but the scientist still needs to instruct the software on which hyper-parameters to explore.
High Bias
Low Variance
Low Bias
High Variance
Prediction Error
Test Sample
Training Sample
Low
High
Model Complexity
underfitting
overfitting
You have 2 main choices to make:
1) the range of values to explore, 2) the actual (and number of) values to explore
Using your best judgment and knowledge of the algorithm and problem, you want hyper-parameters that span the range of low to high model complexity
Range should start here
And end here
range from underfitting to overfitting

Validation Loss metric does not have to be the same as the training loss.

- Sometimes the loss metric for an application is not very easy or even possible to directly minimize.
- We use logistic-loss to find f. but we need a loss metric that enables optimal ranking (such as AUC)
- We might want to minimize training loss. but maximize the validation loss (logistic loss vs AUC)

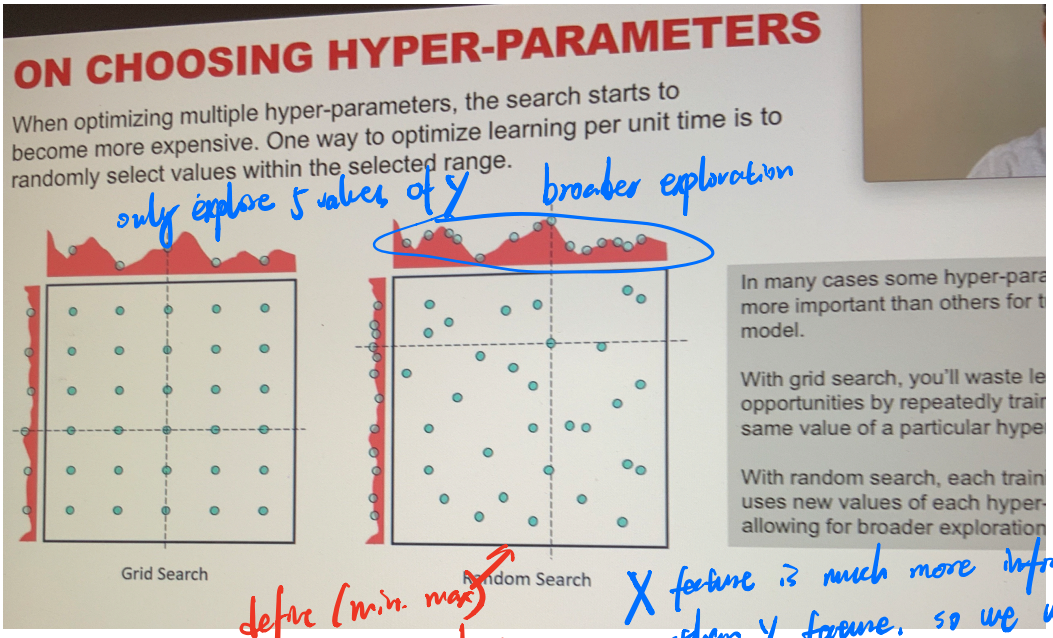
training loss validation loss.

Hyperparameter selection:

1. Experiment with parameters that span the range of low to high model complexity (from underfitting to overfitting).

Grid search,
With grid search. you waste learning opportunities by repeatedly training on the same value of a particular hyper-parameter.

Random Search:
With Random search. each training iteration uses new values of each hyper-parameter, allowing for broader exploration of the space.



Right off the bat, start with simple model (Random Forest or Logistic Regression) to get a baseline.

If there's a significant difference between the performance of Random Forest and Logistic Regression, it means there's a lot of non-linearity in your data that you need to capture.

Iterate towards better Model:

Starting from step 2, try new features and new algorithms. In Real life, simplicity is prefered. so don't use a model that is 5% complexity of your current model, but with 0.% increase in metric.