Statistical learning *

- inputs = predictors, independent variables, features
- output = response, dependent variable

$Y = f(X) + \epsilon$ => f represents the systematic information that X provides about Y
=> ϵ, random error term, independent of X, mean zero

why estimate f? → **Prediction**: we got X, we want Y, we use $\hat{Y} = \hat{f}(X)$ as errors average to 0. $\hat{f}$ treated as blackbox, just want correct Y

Accuracy of $\hat{Y}$ => $E(Y-\hat{Y})^2 = E[f(X)+\epsilon-\hat{f}(X)]^2$

$= [f(X)-\hat{f}(X)]^2 + Var(\epsilon)$

reduced by using a better model, till perfect estimate $\hat{y} = f(x)$ ← reducible + irreducible (lower bound)

reducible → bias^2, Variance

irreducible → The quantity ϵ contain unmeasured variables useful in predicting Y. but also unmeasurable variation

→ **Inference**:
- we need to know the exact form of f.
- better use restrictive model for inference

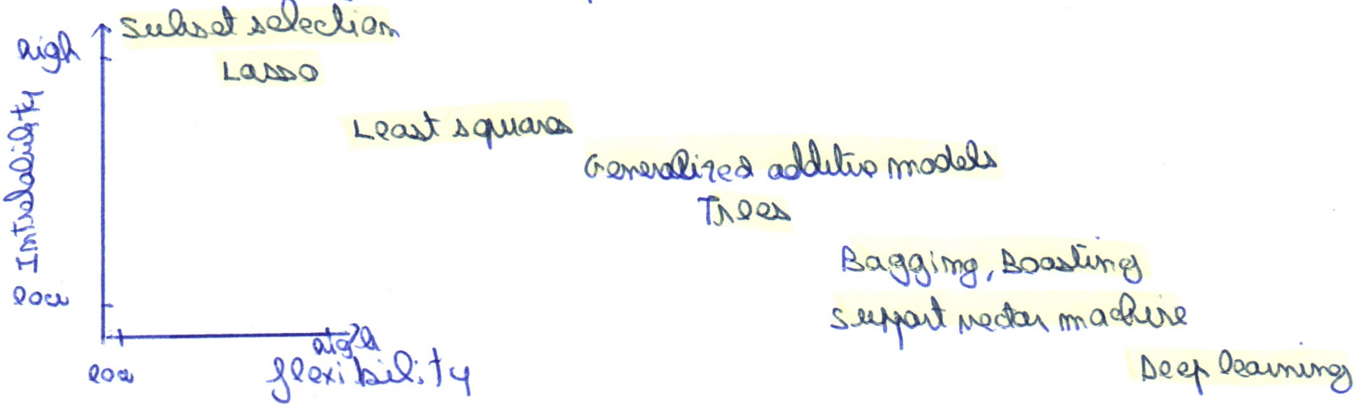
Parametric

-> assumption about the functionnal form or shape of f
-> data fit and train the model

-> usually doesn't match the true unknown f -> so use flexible models that can fit many forms -> but requires more parameters -> which can lead to overfitting.

Non-parametric

-> no explicit assumption -> as close to the data points as possible without being too rough or wiggly. -> need more data.
-> correct amount of smoothness (chs).



Measuring quality of fit (for regression)

- $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$ -> there is no guarantee that the method with the lowest training MSE will also have the lowest test MSE.
-> this is why use cross-validation (chs).
- The expected MSE for a given value x_0, can be decomposed into the sum of

$$E(y_0 - \hat{f}(x_0))^2 = \underbrace{Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2}_{\text{reducible}} + \underbrace{Var(\epsilon)}_{\text{Irreducible}}$$

- we want to reduce variance and biais => both positive => hence Var(e) is the irreductible error.

- variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. In general more flexible methods have higher variance.

- bias refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a simpler model. More flexible => less bias.

classification setting

Same concepts as for MSE in regression but use error rate $\frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$

Bayes classifier

we can reduce error rate, by assigning each observation to the most likely class given its predictors value. $1 - \max_{j} \Pr(Y = j | X = x_0)$ is going to be the test error.

The overall Bayes error is $1 - E(\max_{j} \Pr(Y = j | X))$. it is analogous to the irreductible error.

K-Nearest neighbors

- In theory we would use Bayes, but we don't have the conditional distribution of Y given X.

- Given a positive integer K and a test observation x_0, the KNN classifier first identifies the K points in the training data that are closest to x_0, represented by N_0. It then estimates the conditional probability for class j as the fraction of points in N_0 whose response values equal j: $\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$

- The choice of K has a strong impact. As 1/N ↗ the method become ↗ flexible.

- ↗ flexible => ↘ training error => but test error has a U-shape.