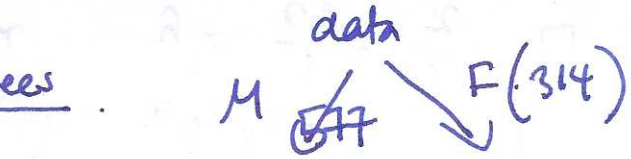
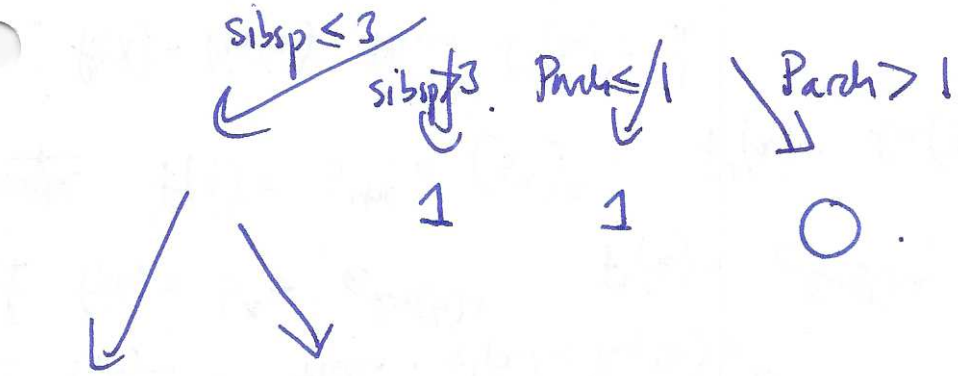
Decision trees





Parch < 4 : 1

Parch ≥ 4 : 0

Parch.

	194	survived	died
0		153	41
1	60	46	14
2	49	30	19
3	4	3	1
4	2	0	2
5	4	1	3
6	1	0	1

---

choose division/outputs to minimize RSS $\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{Rj})$

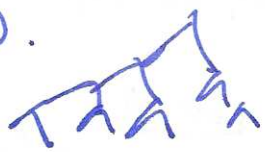
Cost complexity pruning / weakest link pruning

tree $T \subset T_0$

$\alpha = 0$, $T = T_0$

minimize: $\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{Rm})^2 + \alpha |T|$

Fact as α goes from 0 to 1 get nested subtrees.



Assessing model accuracy

training data ⇝ model ⇝ test data.

$X = \{x_i\}$. $\hat{f}(x_i)$.

do MSE here.

mean squared error

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{f}(x_i))^2$$

avoids overfitting: given some data $X = \{x_i\}$.
you can tune any sufficiently complicated model to give 100% accuracy on that data set in such a way it gives bad predictions on test data.

no test data? split training data.

jargon: variance vs bias.

variance: sensitivity of model to changes in data

bias: error due to actual data being nonlinear.

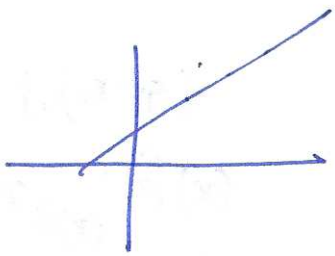
e.g. linear regression very stable.

linear regression — low variance — (maybe) high bias

kNN — higher variance — (maybe) low bias.

multiple linear regression

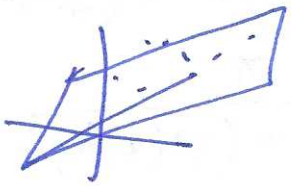
single var regression:



(x_i, y_i).

$y = mx + b$.

multiple vars



$(x_1, x_2, \ldots, x_n)$.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

minimise: $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots \hat{\beta}_p x_{ip})$

$$= \sum_{i=1}^{n} (y_i - [\beta][x]).$$

for predicting z in terms of (x_i, y_i) given best fit plane.

Q: are any of the x_i useful predictors?

- do you need all of them?
- how good is the fit?
- what should we predict? How accurate is the prediction?

---

garden of forking paths.

— everything is probably weakly correlated

— even if things are indep, if looked at 500 regression lines, 1 in 20 things are significant.