Model selection

How do we choose between models M_0 and M_1?

Use M_1 if $p(M_1|\mathcal{D}) > p(M_0|\mathcal{D})$, or

$$\frac{p(M_1|\mathcal{D})}{p(M_0|\mathcal{D})} > 1 .$$

If we assume $p(M_0) = p(M_1) = \frac{1}{2}$,

$$\frac{p(M_1|\mathcal{D})}{p(M_0|\mathcal{D})} = \frac{p(\mathcal{D}|M_1)}{p(\mathcal{D}|M_0)} = \text{Bayes factor (BF)}.$$

Here, $p(\mathcal{D}|M_i)$ = marginal likelihood, or <u>model evidence</u> ($p(\mathcal{D})$ from before)

$BF < \frac{1}{100}$ - decisive evidence for M_0,

$BF > 100$ - decisive evidence for M_1, etc.

<u>Ex.</u> Testing if a coin is fair.

M_0: fair coin, $\theta = 0.5$

M_1: (potentially) biased coin, $0 \leq \theta \leq 1$

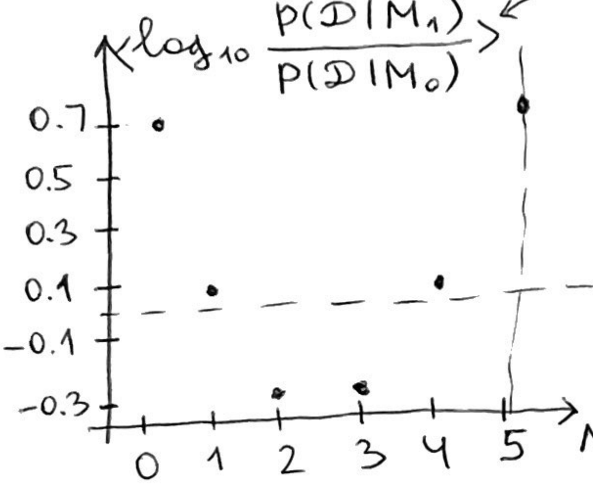
$\mathcal{D}$: $N = 5$ coin tosses

$$\begin{cases} p(\mathcal{D}|M_0) = \left(\frac{1}{2}\right)^N, \\ p(\mathcal{D}|M_1) = \int_0^1 d\theta\, p(\mathcal{D}|\theta)\, \underset{\substack{\uparrow \\ \text{Beta prior}}}{p(\theta)} = \end{cases}$$

$$= \frac{B(\alpha_1+N_1, \alpha_0+N_0)}{B(\alpha_1, \alpha_0)}.$$

Choose $\alpha_1 = \alpha_0 = 1 \Rightarrow$ uniform prior

$N = \underset{\substack{\uparrow \\ \#\text{ tails}}}{N_0} + \underset{\substack{\uparrow \\ \#\text{ heads}}}{N_1}$



$\log_{10} p(\mathcal{D}|M_0) = -N \log_{10}(2) =$
$\simeq -1.505$

$N_1 = 2, 3$ M_0 preferred
$N_1 = 1, 4$ M_1 weakly —"—
$N_1 = 0, 5$ M_1 strongly —"—

Bayesian model selection

How do we choose the best model from a set of models $\mathcal{M}$?

$$\hat{m} = \underset{m \in \mathcal{M}}{\operatorname{argmax}}\, p(m|\mathcal{D}), \quad \text{where}$$

MAP estimate 'under 0-1 loss'

$$p(m \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid m)\, p(m)}{\sum_{m \in M} p(\mathcal{D} \mid m)\, p(m)}.$$

If $p(m) = \frac{1}{\underbrace{|M|}_{\text{size of set } M}}$ $\Leftarrow$ uniform prior,

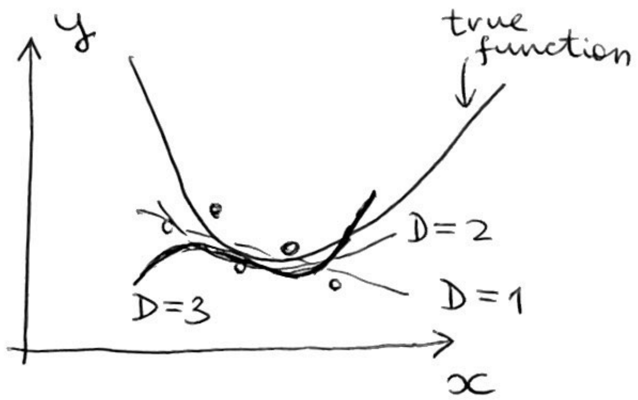
$$\hat{m} = \underset{m \in M}{\operatorname{argmax}}\; p(\mathcal{D} \mid m), \quad \text{where}$$

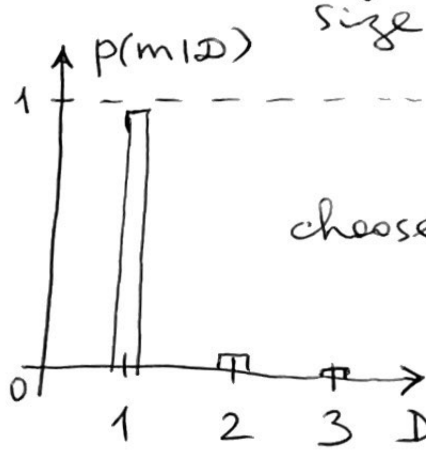
$$\underset{\substack{\text{model}\\ \text{evidence}}}{p(\mathcal{D} \mid m)} = \int d\vec{\theta}\; \underbrace{p(\mathcal{D} \mid \vec{\theta}, m)}_{\substack{\text{favors}\\ \text{more}\\ \text{complex models}}}\; \underbrace{p(\vec{\theta} \mid m)}_{\substack{\text{favors}\\ \text{simpler}\\ \text{models}}}$$

Ex. Polynomial regression

D = polynomial degree, N = sample size

$N=5$





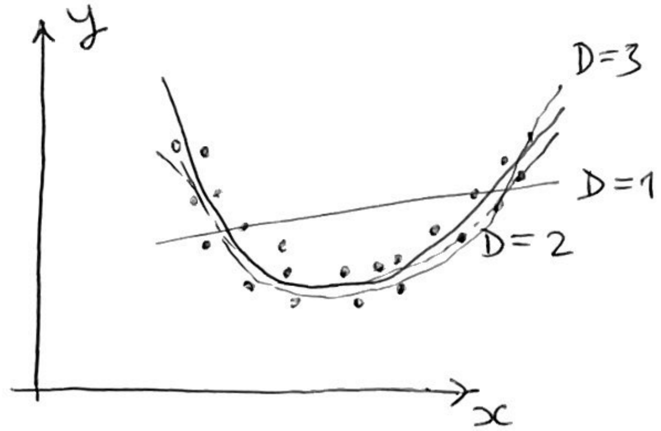
choose $D=1$

Fits correspond to maxima of pred. distribution.

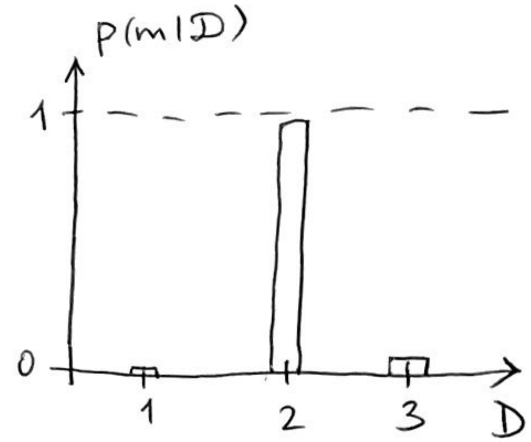
Uniform prior: $p(m) = \frac{1}{3}$.

$$\sum_{m \in M} p(m \mid \mathcal{D}) = 1$$

$N = 30$



Both $D = 2$ & $D = 3$ fit well



choose $D = 2$ → larger sample enables picking a more complex model

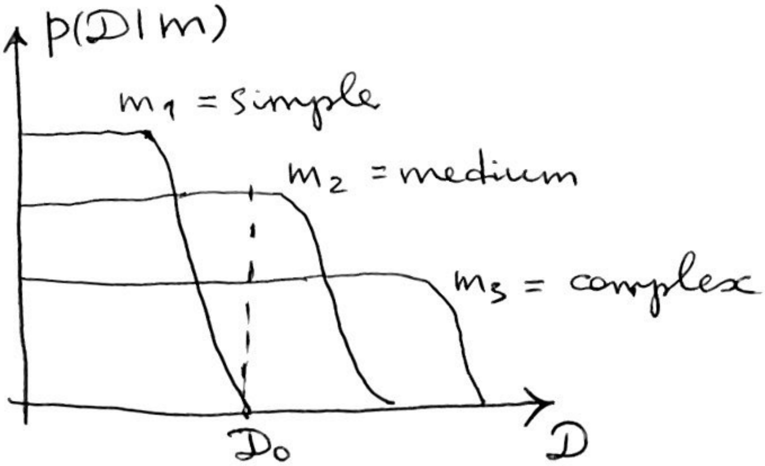
Occam's razor

Consider $\begin{cases} m_1 = \text{simple model} \\ m_2 = \text{complex model} \end{cases}$

If $p(\mathcal{D}|\hat{\theta}_1, m_1) \simeq p(\mathcal{D}|\hat{\theta}_2, m_2)$,

we pick m_1 over m_2 since it is simpler.

Consider summing 'over all possible datasets':
$$\sum_{\mathcal{D}'} p(\mathcal{D}'|m) = 1$$



datasets of increasing complexity

m_2 is chosen b/c it has the highest $p(D_0 | m)$: it can predict D_0 well but does not predict more complex datasets like m_3.