Chapter 6: Frequentist statistics

Bayesian interpretation: Probabilities represent my uncertainty about the world.

Frequentist interpretation: Probabilities represent frequencies of real random outcomes. Properties of the world (θ) are fixed, so they are not random variables.

Me explaining p-value to anyone at all:

p-value

$$\text{pvalue}(\mathcal{D}) \triangleq p(f(\tilde{\mathcal{D}}) \geq f(\mathcal{D}|\tilde{\mathcal{D}} \sim H_0))$$

Confidence interval

Confidence interval:

$$(\ell, u) : P(\ell(\tilde{\mathcal{D}}) \leq \theta \leq u(\tilde{\mathcal{D}}) | \tilde{\mathcal{D}} \sim \theta) = 1 - \alpha$$

Problem 5 (Textbook exercise 3.6): MLE for the Poisson distribution.

The Poisson distribution pmf is defined as

$$\text{Poi}(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!},$$

for $x \in \{0, 1, 2, \dots\}$ where $\lambda > 0$ is the rate parameter. Derive the MLE.

Problem 9 (Textbook exercise 3.20): Comparing naive Bayes with full conditional model

Consider a generative classifier for C classes with class conditional density $p(\mathbf{x}|y)$ and uniform class prior $p(y)$. Suppose all the D features are binary $x_j \in \{0, 1\}$. If we assume all the features are conditionally independent (the naive Bayes assumption), we can write

$$p(\mathbf{x}|y = c) = \prod_{j=1}^{D} \text{Ber}(x_j|\theta_{jc}0$$

This requires DC parameters.

1. Now consider a different model, which we will call the "full" model, in which all the features are fully dependent (i.e. we make no factorization assumptions). How might we represent $p(\mathbf{x}|y = c)$ in this case? How many parameters are needed to represent $p(\mathbf{x}|y = c)$?

2. Suppose we train each model by finding the MLE on a training set of N cases. If the sample size N is very small, which model (naive Bayes or full) is likely to give lower test set error, and why?

3. If the sample size N is very large, which model is likely to give lower test set error, and why?

4. What is the computational complexity (in big-Oh notation) of fitting the full and naive Bayes models respectively as a function of N and D? (You may assume you can convert a D-bit array to an array index in $O(D)$ time.)

5. What is the computational complexity of applying the full and native Bayes models respectively at test time to a single test case?

6. Suppose the test case has missing data. Let $\mathbf{x}_v$ be the visible features of size v and let $\mathbf{x}_h$ be the hidden (missing) features of size h, where $v+h = D$. What is the computational complexity of computing $p(y|\mathbf{x}_v, \hat{\theta}$ for the full and naive bayes models, as a function of v and h?

Problem 7 (Textbook exercise 3.12): MAP estimation for Bernoulli with non-conjugate priors

Suppose we flip a coin N times and observe N_0 tails and N_1 heads. Consider the following prior that believes the coin is either fair or slightly biased towards tails.

$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.4 \\ 0 & \text{otherwise} \end{cases}$$

1. Derive the MAP estimate under this prior.

2. Suppose the true parameter is $\theta = 0.41$. When N is small, will this prior or a Beta$(1, 1)$ prior lead to a MAP estimate that is closer to the true θ? What about when N is large?

Problem 8 (Textbook exercise 3.21). Feature selection using a naive Bayes classifier with binary features.

One way to select features for a naive Bayes classifier is to use mutual information between each feature X_j and the class label Y:

$$I(X_j, Y) = \sum_{x_j} \sum_{j} p(x_j, y) \log \frac{p(x_j, y)}{p(x_j)p(y)}.$$

Let $\pi_c = p(y = c)$, $\theta_{jc} = p(x_j = 1|y = c)$ and $\theta_j = p(x_j = 1) = \sum_c \pi_c \theta_{jc}$. Show that the MI can be computed as follows

$$I(X_j, Y) = \sum_c \left[\theta_{jc} \pi_c \log \frac{\theta_{jc}}{\theta_j} + (1 - \theta_{jc}) \pi_c \log \frac{1 - \theta_{jc}}{1 - \theta_j} \right].$$

Chapter 5: Model selection and making predictions

Utility and loss

Utility: The "goodness" of a decision.

Exercise

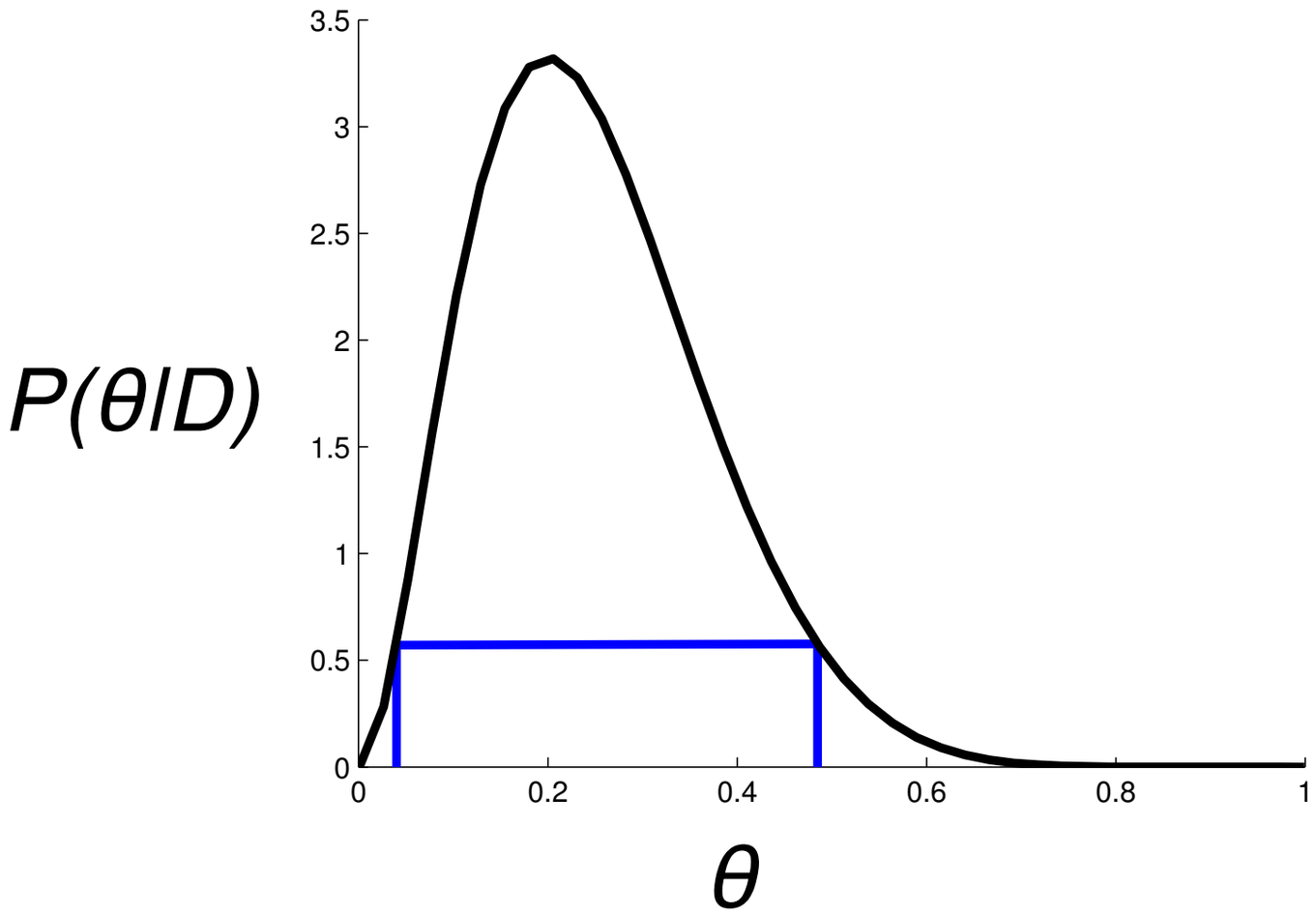
Suppose we estimate the posterior probability of a coin's heads frequency as *p(θ|D).* We need to choose a predicted value of θ, and our loss from this decision will be

$$(\theta^* - \hat{\theta})^2$$

What value of θ should we predict?

Posterior mean, median and mode

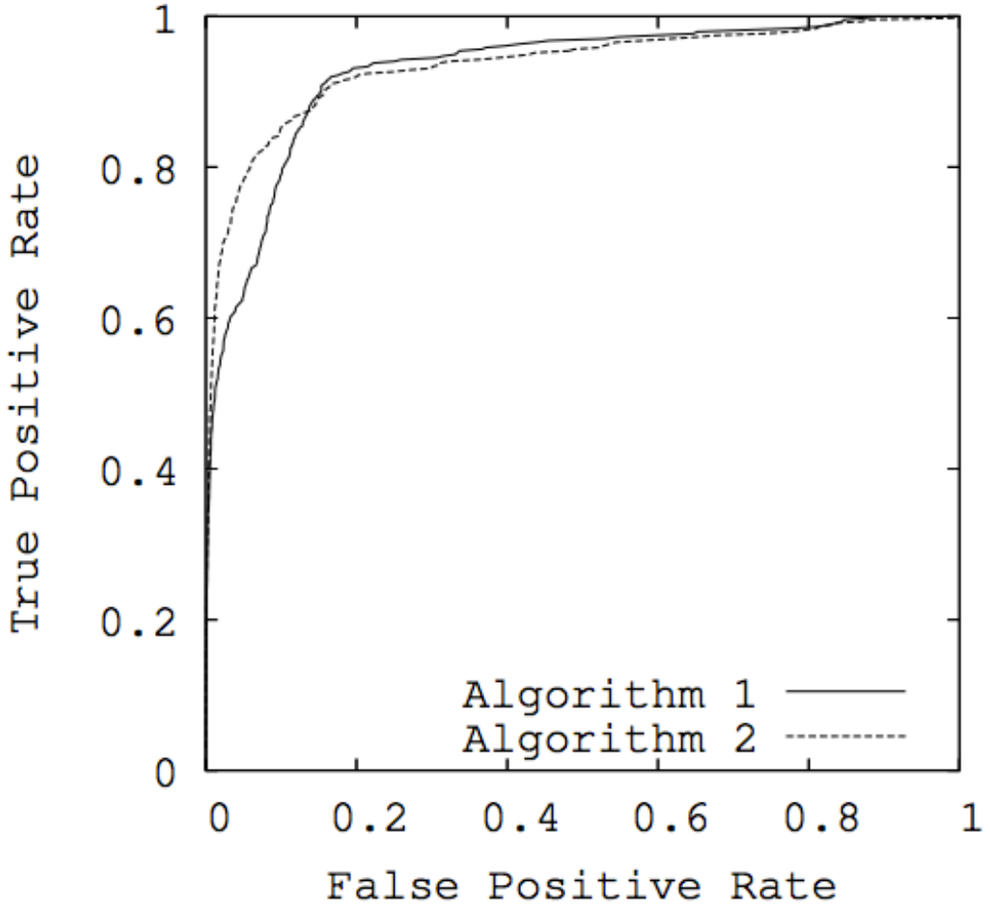
Credible interval

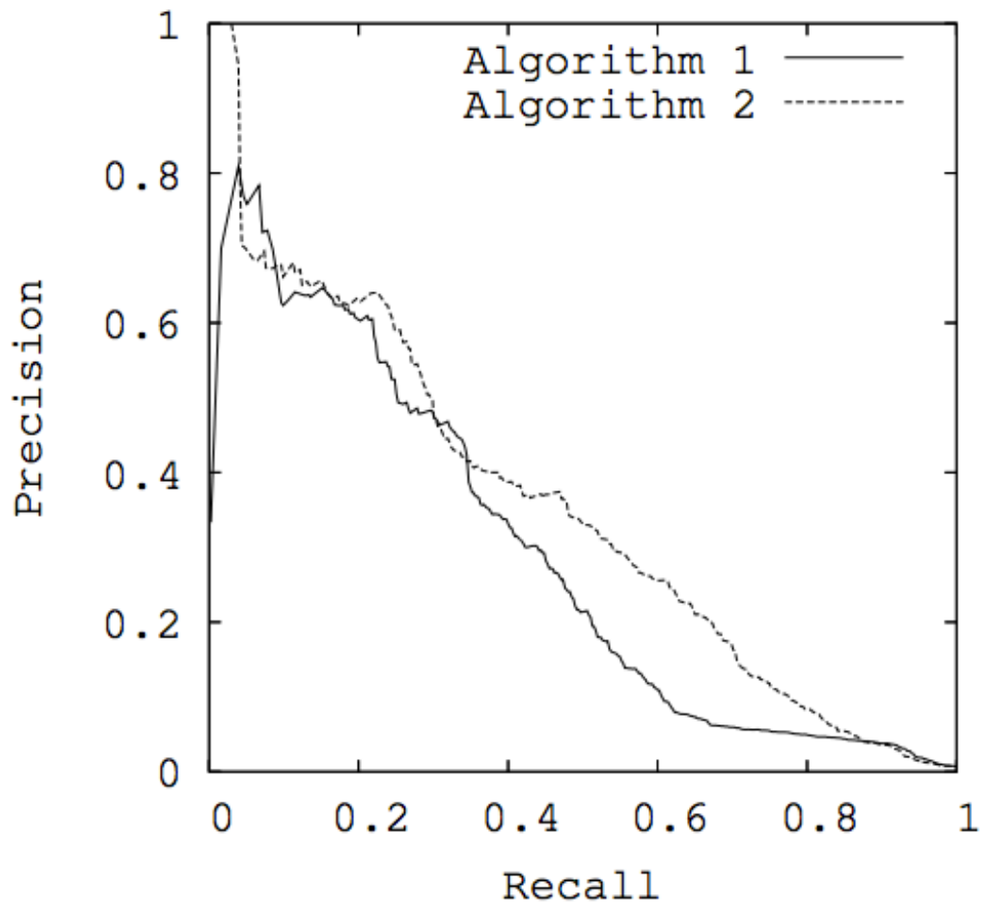


$$C_\alpha(\mathcal{D}) = (\ell, u) : P(\ell \leq \theta \leq u|\mathcal{D}) = 1 - \alpha$$

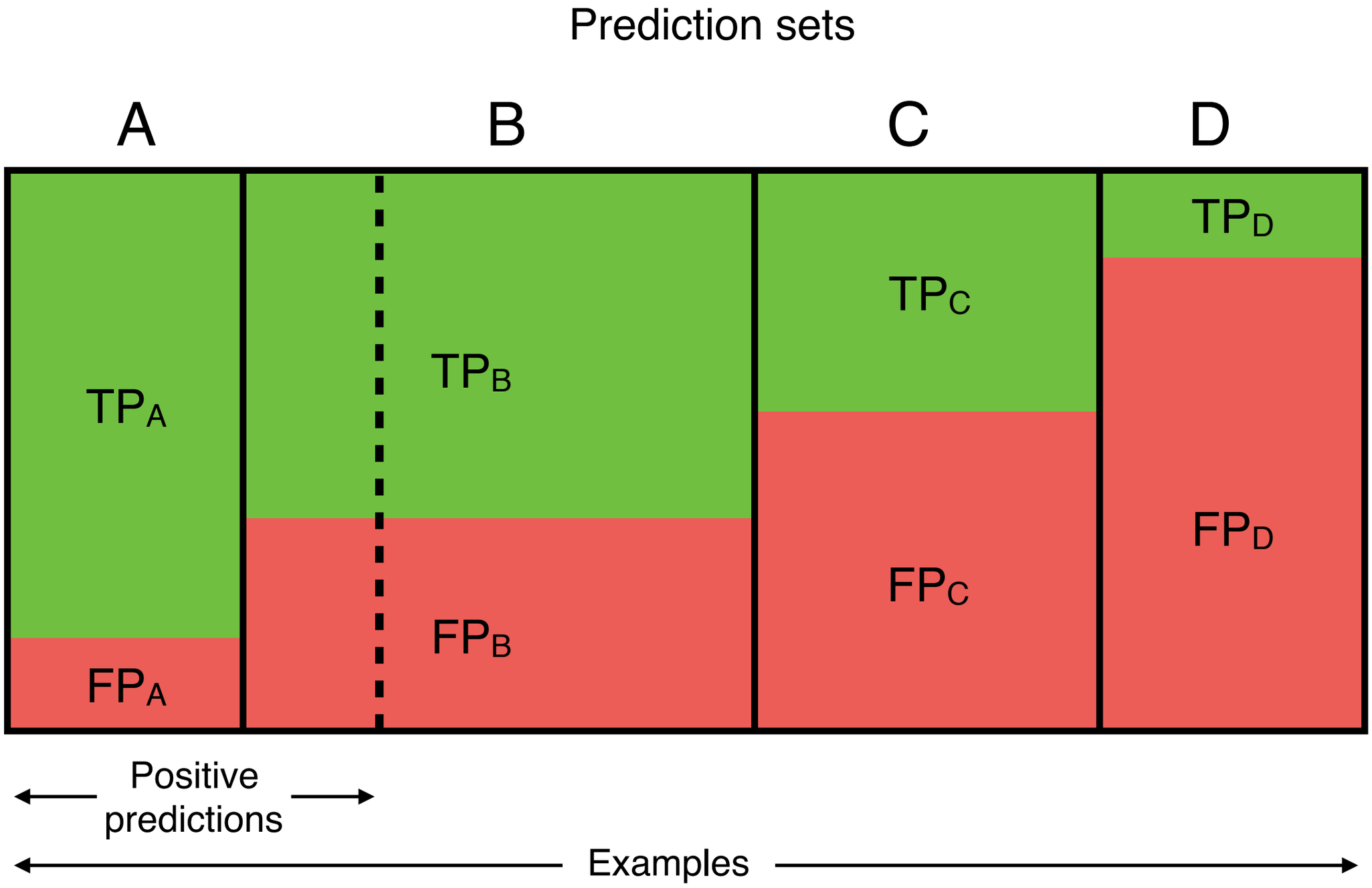
False positive/false negative tradeoff

	$\hat{y}=1$	$\hat{y}=0$
$y=1$	0	L_{FN}
$y=0$	L_{FP}	0

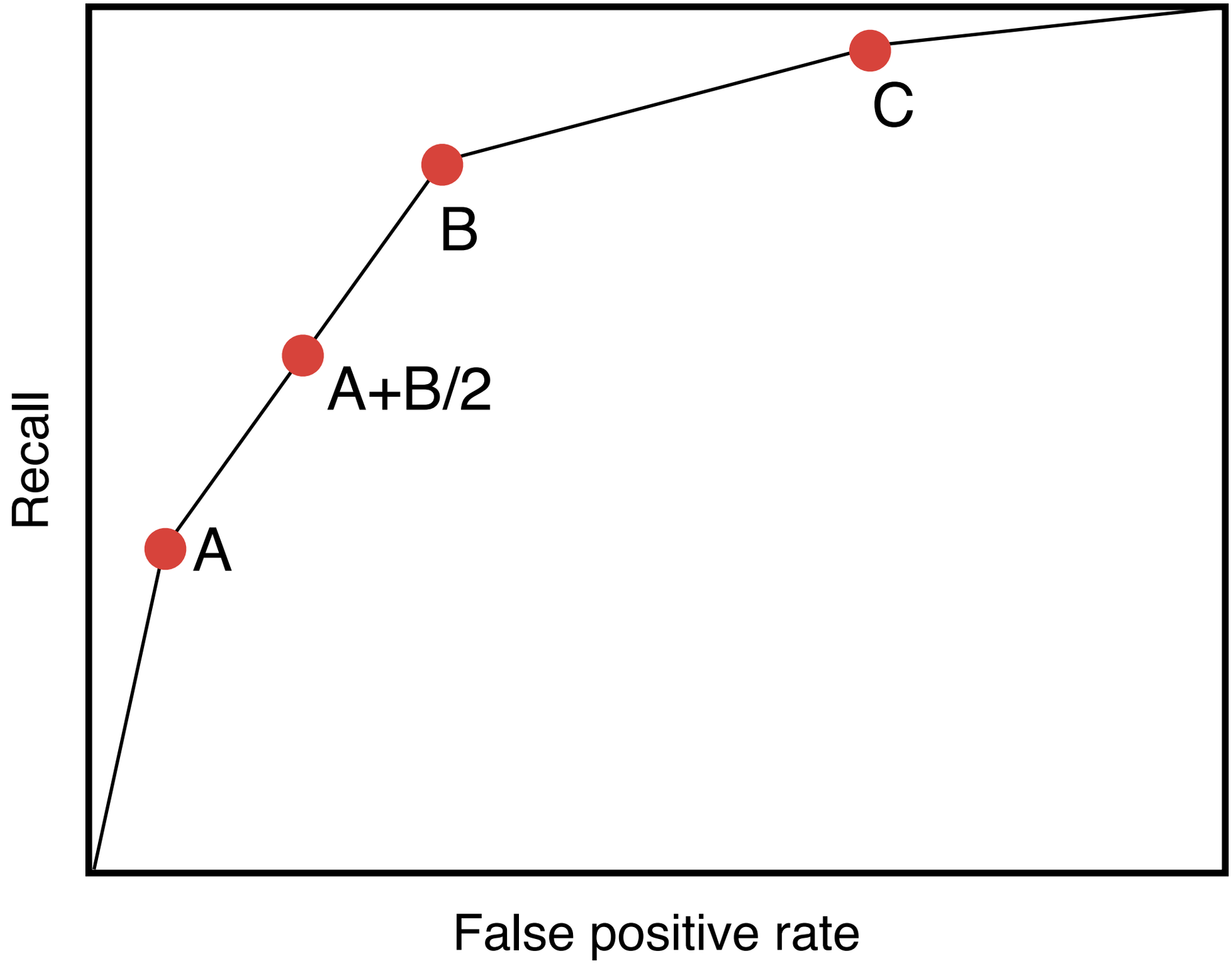




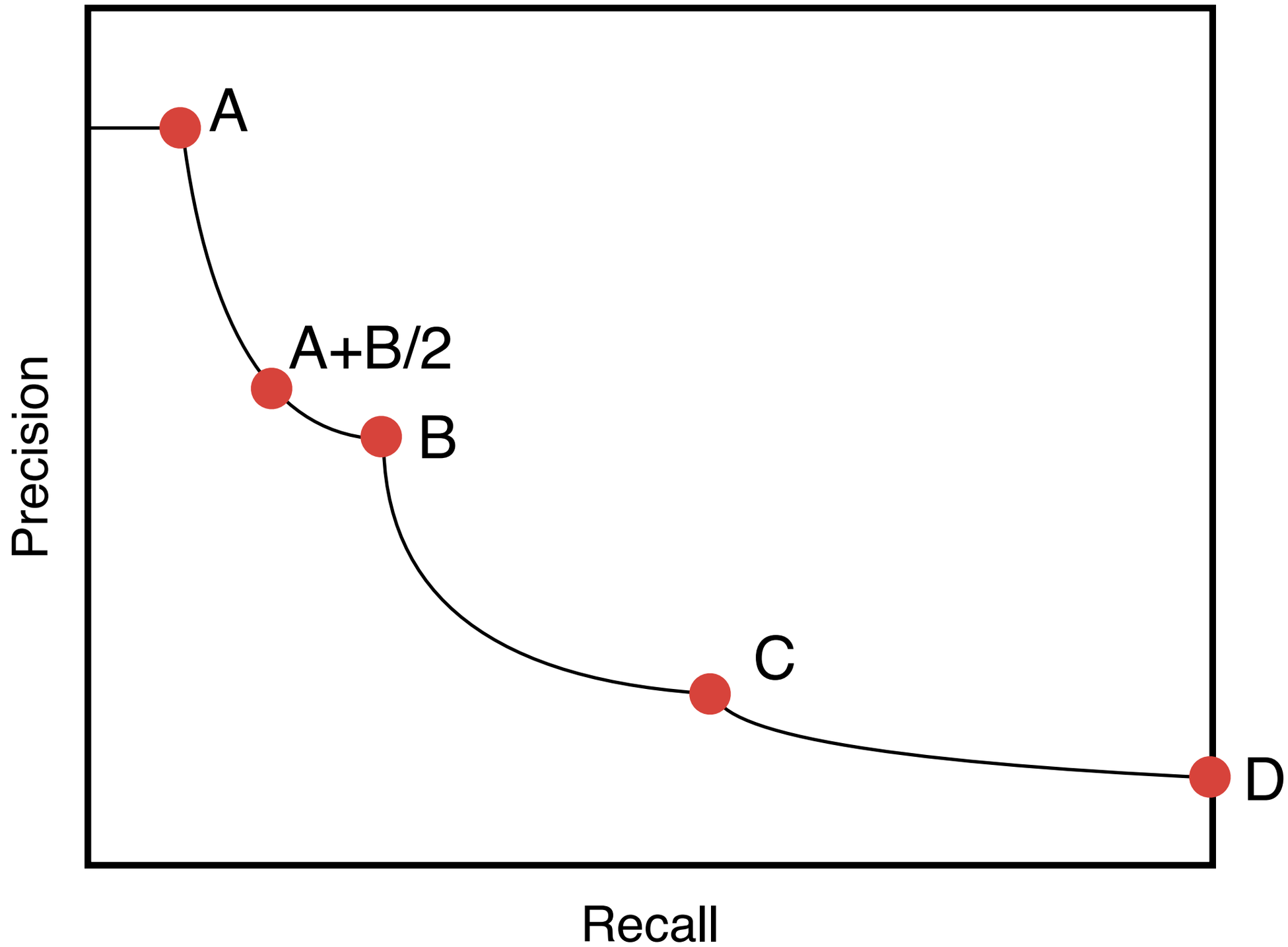
How do you handle ties in prediction probability scores?



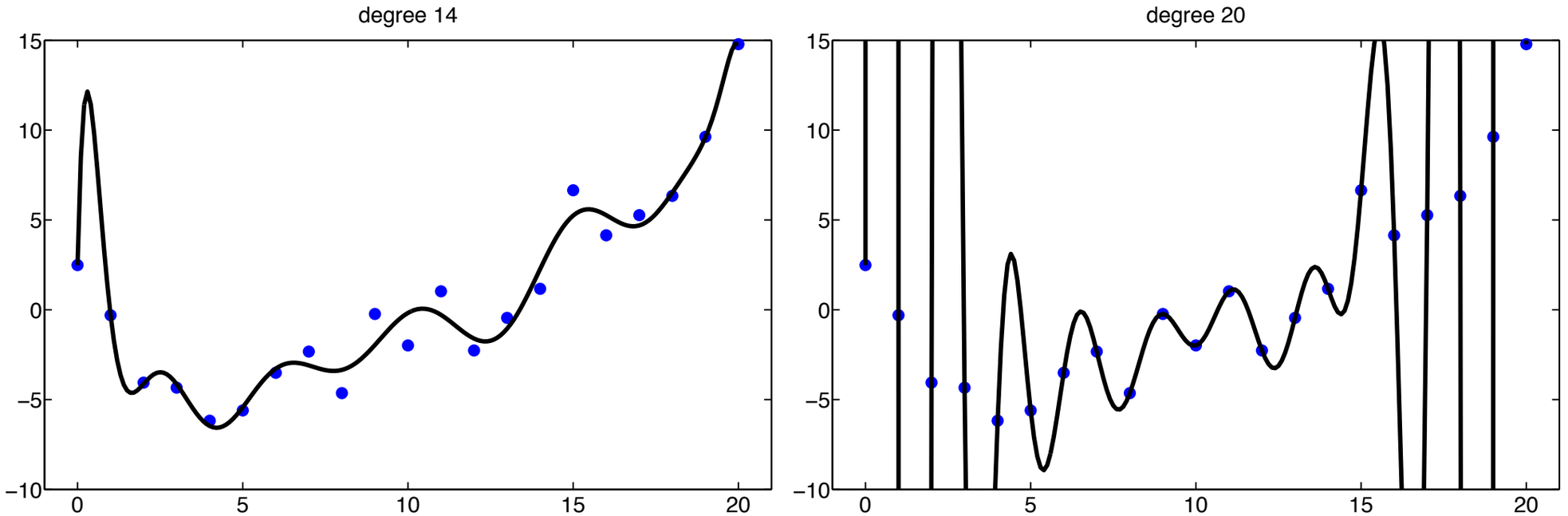
Interpolation in an ROC curve forms a straight line



Interpolation in an PR curve forms convex curve



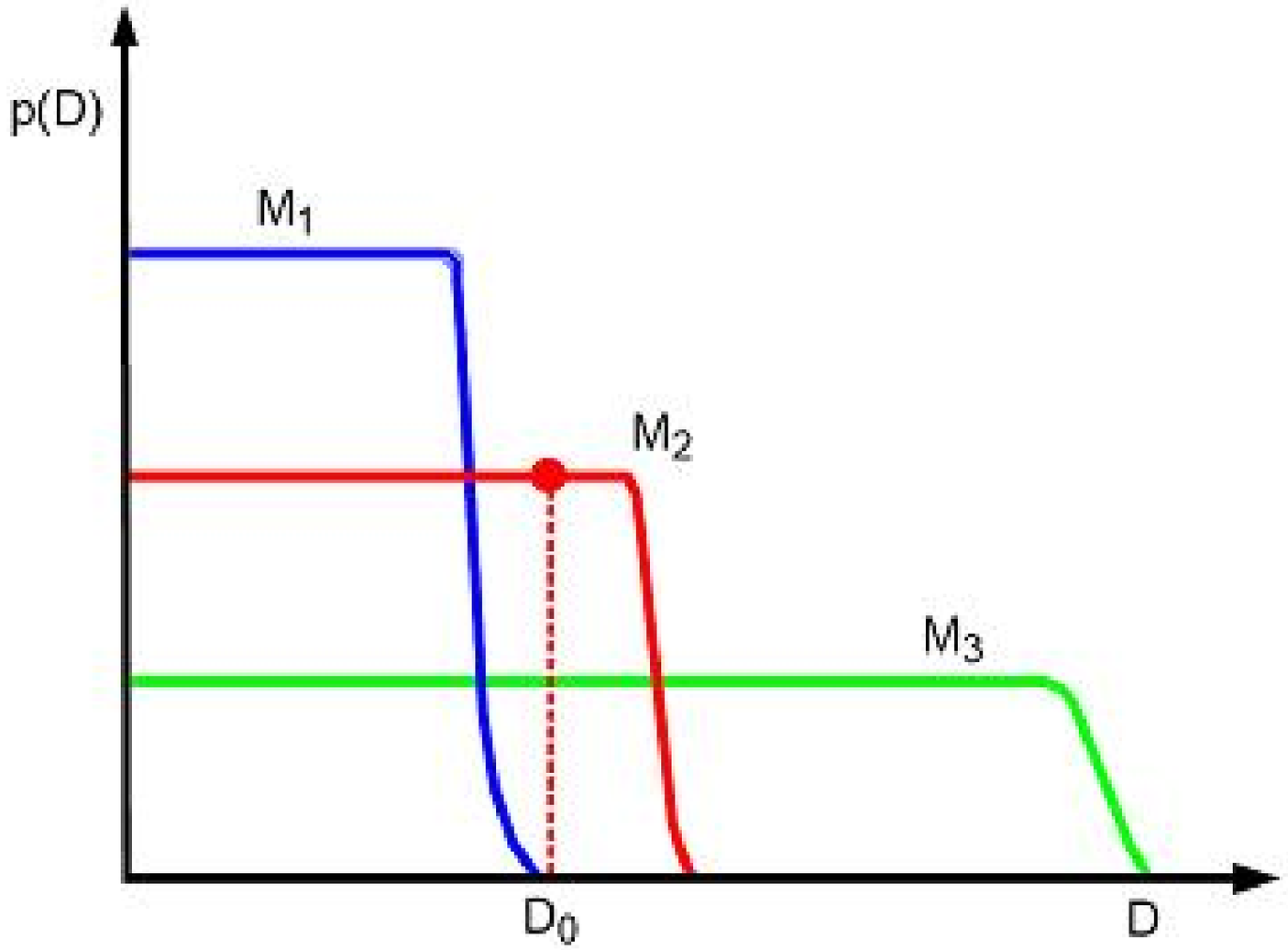
Bayesian model selection



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

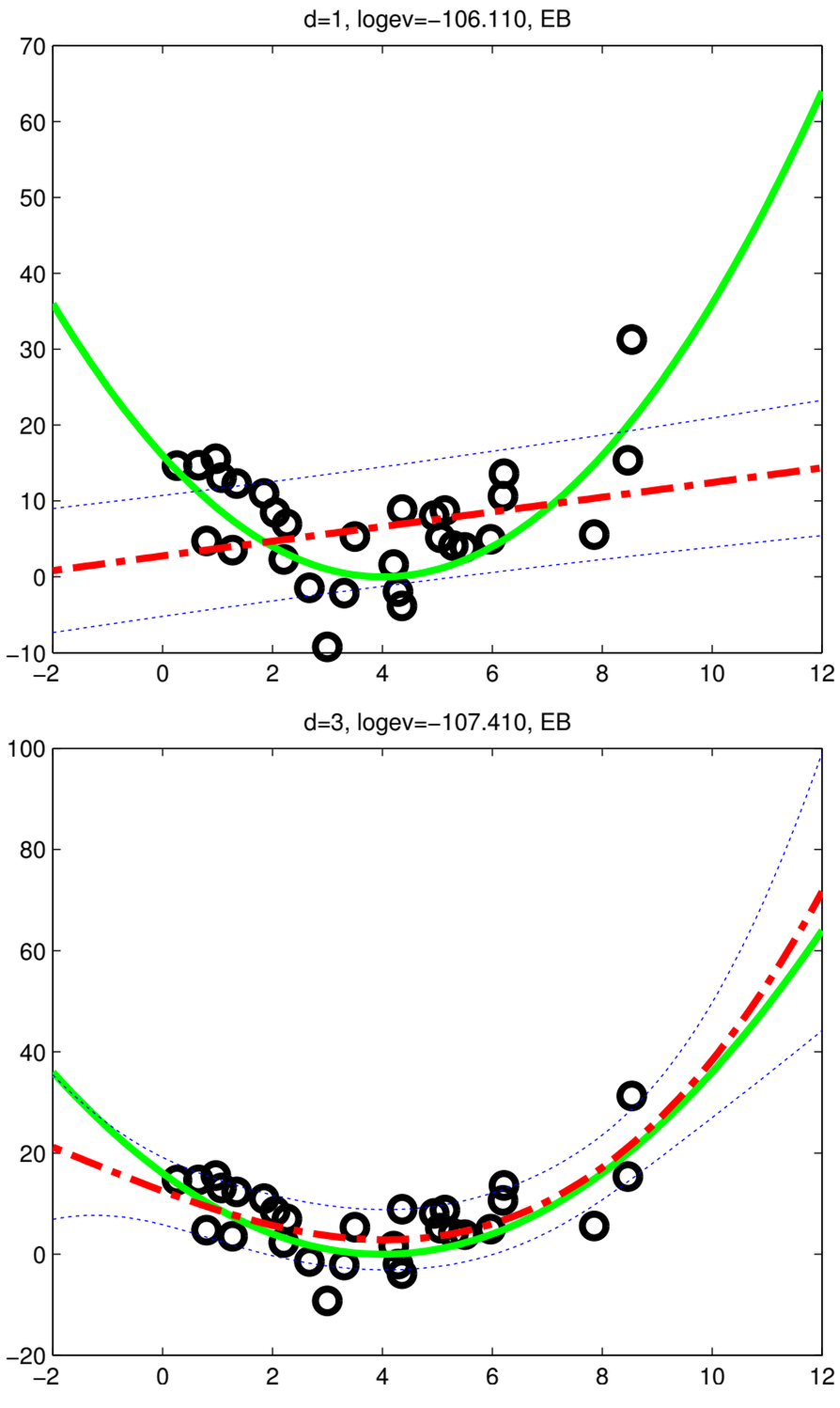
$$p(\mathcal{D}|m) = \int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|m)d\boldsymbol{\theta}$$

Bayesian Occam's razor



"Complexity" of *D*

Bayesian Occam's razor



Problem 5 (Textbook exercise 3.12): MAP estimation for Bernoulli with non-conjugate priors

Suppose we flip a coin N times and observe N_0 tails and N_1 heads. Consider the following prior that believes the coin is either fair or slightly biased towards tails.

$$p(\theta) = \begin{cases} 0.5 & \text{if } \theta = 0.5 \\ 0.5 & \text{if } \theta = 0.4 \\ 0 & \text{otherwise} \end{cases}$$

1. Derive the MAP estimate under this prior.

2. Suppose the true parameter is $\theta = 0.41$. When N is small, will this prior or a Beta$(1, 1)$ prior lead to a MAP estimate that is closer to the true θ? What about when N is large?

Exercise

Suppose we are performing a classification problem where we must predict $y \in \{1, \ldots, C\}$. We have a "reject" option *(C+1)*, where we may refuse to predict any class. We incur the following loss:

$$L(y = j, a = i) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, \ldots, C\} \\ \lambda_r & \text{if } i = C + 1 \\ \lambda_s & \text{otherwise} \end{cases}$$

Given a posterior distribution for *y*, what action should we choose?

Reject option