Exercise 5.8 MLE and model selection for a 2d discrete distribution

(Source: Jaakkola.)

Let x ∈ {0,1} denote the result of a coin toss (x = 0 for tails, x = 1 for heads). The coin is potentially biased, so that heads occurs with probability θ_1. Suppose that someone else observes the coin flip and reports to you the outcome, y. But this person is unreliable and only reports the result correctlyu with probability θ_2; i.e. $p(y|x,\theta_2)$ is given by

	$y=0$	$y=1$
$x=0$	θ_2	$1-\theta_2$
$x=1$	$1-\theta_2$	θ_2

Assume that θ_2 is independent of x and θ_1.

a. Write down the joint probability distribution $p(x,y|\vec{\theta})$ as a 2x2 table, in terms of $\vec{\theta}=(\theta_1,\theta_2)$.

	$y=0$	$y=1$
$x=0$	$(1-\theta_1)\theta_2$	$(1-\theta_1)(1-\theta_2)$
$x=1$	$\theta_1(1-\theta_2)$	$\theta_1\theta_2$

b. Suppose [we] have the following dataset:
   $x=(1,1,0,1,1,0,0)$
   $y=(1,0,0,0,1,0,1)$
   What are the MLEs for θ_1 and θ_2? Justify your answer. Hint: note that the likelihood function factorizes.
   $p(x,y|\vec{\theta})=p(y|x,\theta_2)p(x|\theta_1)$
   Since $log\left(p(x,y|\vec{\theta})\right)=\sum_{i=1}^{n} log(p(y|x,\theta_2))+\sum_{i=1}^{n} log(p(x|\theta_1))$, we can maximize the terms independently.

For n independent trials, where the probability of success for a trial is given by θ...

$$probability(Data \mid \theta) = \theta^{\left(\sum_{i=1}^{n} x_i\right)} (1-\theta)^{\left(n-\sum_{i=1}^{n} x_i\right)}$$

...so...

$$\log\left(probability(Data \mid \theta)\right) = \log\left(\theta^{\left(\sum_{i=1}^{n} x_i\right)} (1-\theta)^{\left(n-\sum_{i=1}^{n} x_i\right)}\right) = \left(\sum_{i=1}^{n} x_i\right)\log(\theta) + \left(n-\sum_{i=1}^{n} x_i\right)\log(1-\theta)$$

...so...

$$\frac{\partial \log\left(probability(Data \mid \theta)\right)}{\partial \theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n-\sum_{i=1}^{n} x_i}{1-\theta}$$

...setting the partial derivative of the log likelihood equal to 0 and solving for θ...

$$\frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n-\sum_{i=1}^{n} x_i}{1-\theta} = 0$$

$$\frac{\sum_{i=1}^{n} x_i}{\theta} = \frac{n-\sum_{i=1}^{n} x_i}{1-\theta}$$

$$\theta(1-\theta)\frac{\sum_{i=1}^{n} x_i}{\theta} = \theta(1-\theta)\frac{n-\sum_{i=1}^{n} x_i}{1-\theta}$$

$$(1-\theta)\sum_{i=1}^{n} x_i = \theta\left(n-\sum_{i=1}^{n} x_i\right)$$

$$\sum_{i=1}^{n} x_i - \theta\sum_{i=1}^{n} x_i = \theta n - \theta\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i = \theta n$$

$$\theta = \frac{\sum_{i=1}^{n} x_i}{n}$$

θ_1 = proportion of x values that are 1 = 4 / 7

θ_2 = proportion of y values that equal the corresponding x values = 4 / 7

What is $p(D|\hat{\boldsymbol{\theta}}, M_2)$ where M_2 denotes this 2-parameter model? (You may leave your answer in fractional form if you wish)

$$p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_2) = \left(\frac{4}{7}\right)^4 \left(1-\frac{4}{7}\right)^{(7-4)} \left(\frac{4}{7}\right)^4 \left(1-\frac{4}{7}\right)^{(7-4)} = \left(\frac{4}{7}\right)^8 \left(\frac{3}{7}\right)^6 = \frac{47{,}775{,}744}{678{,}223{,}072{,}849} \approx 0.0000704$$

c. Now consider a model with 4 parameters, $\vec{\theta} = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})$, representing $p(x, y|\vec{\theta}) = \theta_{x,y}$. (Only 3 of these parameters are free to vary, since they must sum to one.) What is the MLE of $\vec{\theta}$? What is $p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_4)$ where M_4 denotes this 4-parameter model?

$$\vec{\theta} = \left(\frac{2}{7}, \frac{1}{7}, \frac{2}{7}, \frac{2}{7}\right)$$

$$p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_4) = \left(\frac{2}{7}\right)^2 \left(\frac{1}{7}\right)^1 \left(\frac{2}{7}\right)^2 \left(\frac{2}{7}\right)^2 = \left(\frac{2}{7}\right)^6 \frac{1}{7} = \frac{64}{823{,}543} \approx 0.0000777$$

d. Suppose we are not sure which model is correct. We compute the leave-one-out cross validated log likelihood of the 2-parameter model and the 4-parameter model as follows:

$$L(m) = \sum_{i=1}^{n} log\left(p\left(x_i, y_i | m, \widehat{\boldsymbol{\theta}}(\mathcal{D}_{-i})\right)\right)$$

and $\widehat{\boldsymbol{\theta}}(\mathcal{D}_{-i})$ denotes the MLE computed on $\mathcal{D}$ excluding row i. Which model will CV pick and why?

Hint: notice how the table of counts changes when you omit each training case one at a time.

For the 2-parameter model, we have …

(log(3/6)+log(3/6)) + (log(3/6)+log(1-4/6)) + (log(1-4/6)+log(3/6)) + (log(3/6)+log(1-4/6)) + (log(3/6)+log(3/6)) + (log(1-4/6)+log(3/6)) + (log(1-4/6)+log(1-4/6)) ≈ -12.14

For the 4-parameter model, we have …

log(1/6) + log(1/6) + log(1/6) + log(1/6) + log(1/6) + log(1/6) + log(0/6) ≈ -10.75 - ∞ = - ∞

[There's a harsh penalty for saying the observed data could never happen.]

Nota Bene: if we use the training data to report risk, the more complex model wins [not a surprise; we do **not** use the training data for model selection].

e. Recall that an alternative to CV is to use the BIC score, defined as

$$BIC(\mathcal{M}, \mathcal{D}) \triangleq log\left(p\left(\mathcal{D} | \widehat{\boldsymbol{\theta}}_{MLE}\right)\right) - \frac{dof(\mathcal{M})}{2}\log(N)$$

where $dof(\mathcal{M})$ is the number of free parameters in the model. Compute the BIC scores for both models (use log base e). Which model does BIC prefer?

For the 2-parameter model we have …

$$log\left(\frac{47{,}775{,}744}{678{,}223{,}072{,}849}\right) - \frac{2}{2}log(7) \approx -11.51$$

For the 4-parameter model we have …

$$log\left(\frac{64}{823{,}543}\right) - \frac{3}{2}log(7) \approx -12.38$$

Both Leave-One-Out Cross Validation (LOOCV) and the Bayesian Information Criterion (BIC) prefer the simpler 2-parameter model [values closer to zero are preferred].