Statistical Rethinking: Chapter 6 - Overfitting, Regularization, and Information Criteria

Chris Hayduk

February 3, 2019

1 Easy

Problem 6E1.
State the three motivating criteria that define information entropy.

The three motivating criteria that define information entropy are:

1. The measure of uncertainty should be continuous. If it were not, then an arbitrarily small change in any of the probabilities would result in a massive change in uncertainty.

2. The measure of uncertainty should increase as the number of possible events increases. For example, suppose there are two cities that need weather forecasts. In the first city, it rains on half of the days in the year and is sunny on the others. In the second, it rains, shines, and hails, each on 1 out of every 3 days in the year. We'd like our measure of uncertainty to be larger in the second city, where there is one more kind of event to predict.

3. The measure of uncertainty should be additive. What this means is that if we first measure the uncertainty about rain or shine (2 possible events) and then the uncertainty about hot or cold (2 different possible events), the uncertainty over the four combinations of these events - rain/hot, rain/cold, shine/hot, shine/cold - should be the sum of the separate uncertainties.

Problem 6E2.
Suppose a coin is weighted such that, when it is tossed and lands on a table, it comes up heads 70% of the time. What is the entropy of this coin?

$$H(p) = -Elog(p_i) = -\sum_{i=1}^{n} p_i log(p_i) = -(0.7 * log(0.7) + 0.3 * log(0.3)) \approx 0.610864$$

Problem 6E3.
Suppose a four-sided die is loaded such that, when tossed onto a table, it shows "1" 20%, "2" 25%, "3" 25%, and "4" 30% of the time. What is the entropy of this die?

$$H(p) = -Elog(p_i) = -\sum_{i=1}^{n} p_i log(p_i)$$
$$= -(0.2 * log(0.2) + 0.25 * log(0.25) + 0.25 * log(0.25) + 0.3 * log(0.3)) \approx 1.37623$$

Problem 6E4.
Suppose another four-sided die is loaded such that it never shows "4". The other three sides show equally often. What is the entropy of this die?

$$H(p) = -Elog(p_i) = -\sum_{i=1}^{n} p_i log(p_i)$$
$$= -(0.3333 * log(0.3333) + 0.3333 * log(0.3333) + 0.3333 * log(0.3333)) \approx 1.09498$$

2 Medium

Problem 6M1.
Write down and compare the definitions of AIC, DIC, and WAIC. Which of these criteria is most general? Which assumptions are required to transform a more general criterion into a less general one?

Information criteria:
AIC = $\mathrm{D}_t rain + 2p$
$DIC = \overline{D} + (\overline{D} - \hat{D}) = \overline{D} + p_D$
$WAIC = -2(lppd - p_W AIC)$
WAIC is the most general, as it does not require a multivariate Guassian posterior as in DIC AIC, and it does not require uninformative priors as in AIC.

Problem 6M2.
Explain the difference between model selection and model averaging. What information is lost under model selection? What information is lost under model averaging?

Model selection involves using DIC/WAIC in combination with the estimates and posterior predictive checks from each model in order to choose the "best" model out of a given set. Model averaging means using DIC/WAIC to construct a posterior predictive distribution that exploits what we know about relative accuracy of the models. This helps guard against overconfidence in model structure, in the same way that using the entire posterior distrbution helps guard against overconfidence in parameter values.

Model selection causes us to lose the information provided by the other models in the set that we're considering. Model averaging causes us to have a much more conservative (ie. wider) interval for μ.

Problem 6M3.
When comparing models with an information criterion, why must all models be fit to exactly the same observations? What would happen to the information criterion values if the models were fit to different numbers of observations? Perform some experiments if you are not sure.

The model fit to fewer observations will almost always have a better deviance and AIC/DIC/WAIC value because it has been asked to predict less.

Problem 6M4.
What happens to the effective number of parameters, as measured by DIC or WAIC, as a prior becomes more concentrated? Why? Perform some experiments if you are not sure.

Regularizing priors constrain a model's flexibility. Since the effective number of parameters measures how flexible the model is, as a prior becomes more concentrated, it reduces the effective number of parameters.

Problem 6M5.
Provide an informal explanation of why informative priors reduce overfitting.

Informative priors are generally very narrow around their mean. That is, they tend to have small standard deviations. Since the machine will be very skeptical of values that are more than two standard deviations above or below the prior's mean, a small standard deviation will restrict the paramter to values close to the prior's mean.

Problem 6M6.
Provide an informal explanation of why overly informative priors result in underfitting.

An overly informative prior will have an excessively tight prior distribution. As a result, the parameter may be constrained to a range that does not contain its "true" value, thus causing the model to underfit the data.

3 Hard

All practice problems to follow use the same data. Pull out the old Howell !Kung demography data and split it into two equally sized data frames. Here's the code to do it:

```
data(Howell1)

d <- Howell1

d$age <- (d$age - mean(d$age))/sd(d$age)

set.seed(1000)

i <- sample(1:nrow(d), size=nrow(d)/2)
```

```
d1 <- d[i, ]

d2 <- d[-i, ]
```

You now have two randomly formed data frames, each with 272 rows. The notion here is to use the cases in d1 to fit models and the cases in d2 to evaluate them. The set.seed command just ensures that everyone works with the same randomly shuffled data.

Now let h_i and x_i be the height and centered age values, respectively, on row i. Fit the folowing models to the data in d1:

- M_1: $h_i \sim \text{Normal}(\mu_i, \sigma)$
  $\mu_i = \alpha + \beta_1 x_i$
- M_2: $h_i \sim \text{Normal}(\mu_i, \sigma)$
  $\mu_i = \alpha + \beta_1 x_i + \beta_2 {x_i}^2$
- M_2: $h_i \sim \text{Normal}(\mu_i, \sigma)$
  $\mu_i = \alpha + \beta_1 x_i + \beta_2 {x_i}^2 + \beta_3 {x_i}^3$
- M_3: $h_i \sim \text{Normal}(\mu_i, \sigma)$
  $\mu_i = \alpha + \beta_1 x_i + \beta_2 {x_i}^2 + \beta_3 {x_i}^3$
- M_4: $h_i \sim \text{Normal}(\mu_i, \sigma)$
  $\mu_i = \alpha + \beta_1 x_i + \beta_2 {x_i}^2 + \beta_3 {x_i}^3 + \beta_4 {x_i}^4$
- M_5: $h_i \sim \text{Normal}(\mu_i, \sigma)$
  $\mu_i = \alpha + \beta_1 x_i + \beta_2 {x_i}^2 + \beta_3 {x_i}^3 + \beta_4 {x_i}^4 + \beta_5 {x_i}^5$
- M_6: $h_i \sim \text{Normal}(\mu_i, \sigma)$
  $\mu_i = \alpha + \beta_1 x_i + \beta_2 {x_i}^2 + \beta_3 {x_i}^3 + \beta_4 {x_i}^4 + \beta_5 {x_i}^5 + \beta_6 {x_i}^6$

Use map to fit these. Use weakly regularizing priors for all parameters.

Note that fitting all of these polynomials to the height-by-age relationship is not a good way to derive insight. It would be better to have a simpler approach that would allow for more insight, like perhaps a piecewise linear model. But the set of polynomial families above will serve to help you practice and understand model comparison and averaging.