Chapter 6

stevenjin8

August 5, 2021

Exercises

Exercise 1

In this question we compare the misclassification rate of an arbitrary classifier on a random dataset versus the misclassification rate using leave one out cross validation (LOOCV).

Since we have a completely random dataset (i.e. $\mathbf{x}_i$ does not help us predict y_i) and the classes are evenly distributed ($N_1 = N_2$), the lowest misclassification rate any classifier could achieve would be 0.5. We do not consider the classifier getting lucky.

Now consider the misclassification rate using LOOCV:

$$\text{misclassification} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f_m^{-i}(\mathbf{x}_i)),$$

where L is a $0-1$ loss and $f_m^{-i}(\mathbf{x}_i)$ is the predicted value of y_i given $\mathbf{x}_i$ and $\mathcal{D}_{-i}$. Since the data is random the best performing classifer will be the one that always picks the most common label.

Taking a closer look at $\mathcal{D}_{-i}$, if $y_i = 1$ then there will be $N_1 - 1$ examples of label 1 and N_2 examples of label 2 in $\mathcal{D}_i$. Since there are more examples of label 2, the best classifier will be the one that always picks $y = 2$ and will misclassify y_i. If $y_i = 2$ then, by the same logic, the best classifier trained on $\mathcal{D}_i$ will always predict $y = 1$ and will also misclassify y_i. Since our model's prediction on y_i given training data $\mathcal{D}_i$ and $\mathbf{x}_i$ (not that knowing $\mathbf{x}_i$ makes a difference) will always be wrong, our LOOCV misclassification rate is 0, which is quite far from our earlier 0.5.

From this exercise we see an extreme example of LOOCV being pessimistic. This pessimism is not surprising. As long as data is quality, the more of it the better. Whenever we do some sort of cross validation, we set aside some data to evaluate the performance on the model and train the model using the rest of the data (in LOOCV the training set has one example). It follows that the model trained on on the test set will not be as good as the one trained on all the data. Further, since we are doing

the test-evaluate process on every data point, it is unlikely that the model will luck out with an easy test set. It is no surprise then that LOOCV is consistently pessimistic.

Another way to view the disparity between the two misclassification rates is with the composition of the test set. We want the test set to be representative of the data. This usually comes in the form of ensuring that the test set has the same label distribution as the data. Unfortunately, this is not possible with LOOCV because the test set has only element (unless there is only one class but why even bother with a model then). When the model relies heavily on the prior distributions of the classes, rather than the feature vectors, the effects the effects of an unrepresentative test set can be best seen, as in this exercise.

Exercise 2

In this exercise, we fit a normal prior to some data. Each data point comes from a separate normal distribution with a fixed and shared variance and a mean that follows the prior distribution.

a. Recall that for some model with parameters $\boldsymbol{\theta}$ and hyperparameters $\boldsymbol{\eta}$, the ML-II estimate is given by

$$\begin{aligned}\hat{\boldsymbol{\eta}} &= \underset{\boldsymbol{\eta}}{\operatorname{argmax}} \left[\log \int p(\mathcal{D}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\eta}) d\boldsymbol{\theta} \right] \\ &= \underset{\boldsymbol{\eta}}{\operatorname{argmax}} \left[\log p(\mathcal{D}|\boldsymbol{\eta}) \right].\end{aligned}$$

Next, we find the log-likelihood in terms of m_0 and τ_0^2. From the results of section 5.6.2.2, we have

$$\begin{aligned}\hat{m}_0 &= \frac{1}{N} \sum_{i=0}^{6} Y_i \\ &= 1527.5\end{aligned}$$

, and

$$\begin{aligned}\hat{\tau}_0^2 &= \max(0, s^2 - \sigma^2) \\ &= 1754.3\end{aligned}$$

, where $s^2 = 2254.3$ is the empirical variance and $\sigma^2 = 500$.

b. From section 5.6.2

$$\begin{aligned}\mathbb{E}[\theta_1|\mathcal{D}, \tau_0^2, m_0] &= \frac{\tau_0^2 Y_1 + \sigma^2 m_0}{\tau_0^2 + \sigma^2} \\ &= 1514\end{aligned}$$

and

$$\begin{aligned}\text{var}[\theta_1|\mathcal{D}, \tau_0^2, m_0] &= \frac{\tau_0^2\sigma^2}{\tau_0^2 + \sigma^2} \\ &= 389.1.\end{aligned}$$

c. The 95% credible interval for θ_i is

$$1514 - 1.96 \cdot 389.1 \leq \theta_i \leq 1514 + 1.96 \cdot 389.1.$$

This interval is probably too thin because we fitted the hyperparameters to the data.

d. If we σ were smaller, our ML-II estimate of m_0 would stay the same since the mean is an unbiased estimator, but τ_0^2 would be much larger since the variance in the data will have had to have come from the prior, not the parameters. Our estimate of $\mathbb{E}[\theta_i|\mathcal{D}, \tau_0^2, m_0]$ would be closer to Y_i since Y_i would be more representative of the θ_1. Finally, $\text{var}[\theta_i|\mathcal{D}, \tau_0^2, m_0]$ would also be smaller as Y_i would be more representative of θ_i.

The most interesting result (personally) is the effect of σ^2 on our estimate of the prior variance τ_0^2. I initially thought that that decreasing σ^2 would also decrease τ_0^2 because the it would leave less room for θ_i to vary. But the fact is the opposite if we think of the variance in a data as a sum of the prior variance τ_0^2 and the variance σ^2.

Exercise 3

We use the alternate formula for variance and the fact that $\mathbb{E}[X^2] > \mathbb{E}[X]^2$ for this proof:

$$\begin{aligned}\mathbb{E}_{X_1,\ldots,X_n \sim \mathcal{N}(|\mu,\sigma^2)}[\hat{\sigma}^2(X_1, \ldots, X_n)] &= \mathbb{E}\left[\frac{1}{n}\sum X_i^2\right] - \mathbb{E}\left[\left(\frac{1}{n}\sum X_i\right)^2\right] \\ &= \sum \mathbb{E}[X_i^2] - \sum_{i,j} \mathbb{E}[X_i X_j] \\ &< \frac{1}{n}\sum \mathbb{E}[X_i^2] - \frac{1}{n}\sum \mathbb{E}[X_i]^2 \\ &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \sigma^2.\end{aligned}$$

We derive the inequality since

$$\begin{aligned}\sum_{i,j} \mathbb{E}[X_i X_j] &= \sum_{i \neq j} \mathbb{E}[X_i]\mathbb{E}[X_j] + \sum_i \mathbb{E}[X_i^2] \\ &> \sum_{i \neq j} \mathbb{E}[X_i]\mathbb{E}[X_j] + \sum_i \mathbb{E}[X_i]^2 \\ &= \sum_{i,j} \mathbb{E}[X_i]\mathbb{E}[X_j]\end{aligned}$$

(pay close attention to the indices).

Exercise 4

We want to find

$$\hat{\sigma}^2 = \underset{\sigma^2}{\operatorname{argmax}} \log p(\mathcal{D}|\mu, \sigma^2).$$

First, we find the log-likelihood in terms of σ^2:

$$\begin{aligned}
\log p(\mathcal{D}|\mu, \sigma^2) &= \log \prod_i \frac{1}{\sqrt{\pi 2\sigma^2}} \exp\left(\frac{(x_i - \mu)^2}{-2\sigma^2}\right) \\
&= \sum_i \frac{(x_i - \mu)^2}{-2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2).
\end{aligned}$$

Next, we derive with respect to σ^2 and set equal to zero:

$$\begin{aligned}
\frac{\partial}{\partial \sigma^2} \log p(\mathcal{D}|\mu, \sigma^2) &= 0 \\
\sum_i \frac{\partial}{\partial \sigma^2}\left[\frac{(x_i - \mu)^2}{-2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right] &= 0 \\
\sum_i \frac{(x_i - \mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} &= 0 \\
\sum_i \frac{(x_i - \mu)^2}{2\sigma^4} &= \sum_i \frac{1}{2\sigma^2} \\
\frac{1}{n}\sum_i (x_i - \mu)^2 &= \hat{\sigma}^2.
\end{aligned}$$

This estimate for σ^2 is unbiased because

$$\begin{aligned}
\mathbb{E}_{X_1,\ldots,X_n \sim \mathcal{N}(\mu,\sigma^2)}[\hat{\sigma}^2(X_1, \ldots, X_n)] &= \mathbb{E}\left[\frac{1}{n}\sum X_i^2\right] - \mu^2 \\
&= \frac{1}{n}\sum_i \mathbb{E}[X^2] - \mu^2 \\
&= \mathbb{E}[X^2] - \mu^2 \\
&= \sigma^2
\end{aligned}$$

where $X \sim \mathcal{N}(\mu, \sigma^2)$.

We see that the sample variance is an unbiased estimator of the true variance when we replace the sample mean with the true mean. This is because observation affects the sample mean causing, the sample variance to be biased. But because observations do not change the true mean, replacing the sample mean with the true mean causes the sample variance to be an unbiased estimator of the true variance.