Optimism of the Training Error

Let us start from the definitions of last time:

$$Err_{\mathcal{T}} = E_{(X_0,Y_0)}\left[L(Y_0, \hat{f}(X_0)) \mid \mathcal{T}\right]$$

Test error, where

- X_0, Y_0 are new (test) point $\rightarrow$ random
- $\mathcal{T}$ is fixed $\quad \mathcal{T} = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$

Averaging over all training sets

$$Err = E_{\mathcal{T}}\left[E_{(X_0,Y_0)}\left[L(Y_0, \hat{f}(X_0)) \mid \mathcal{T}\right]\right]$$

gives the expected error

We also saw the training error

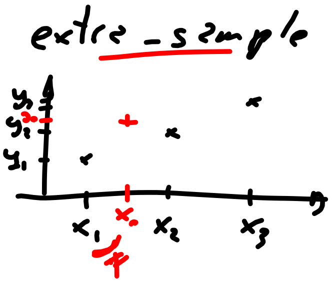
$$\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

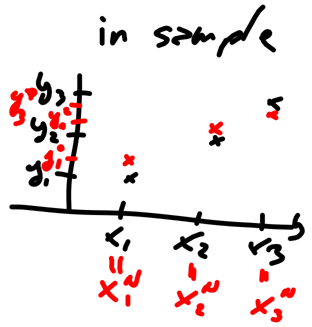
is NOT a good estimate of the test error, because it underestimate it.
Same data used for both training $\hat{f}(x)$ and to test its performance $\rightarrow$ optimistic estimate

The test error can be thought as a extra-sample error (the estimation of the error is computed on new points $x_0 \neq x_i$

We are going to evaluate the optimism of $\overline{err}$ in the in-sample case, i.e. we have new observations in the same points of training set

$$Err_{in} = \frac{1}{N}\sum_{i=1}^{N} E_{Y_0}\left[L(Y_i^0, \hat{f}(x_i)) \mid \mathcal{T}\right]$$





- only Y_0 is random, new y_i^0 for $x_1, \dots, x_N$

Definition: we define optimism the difference between Err_{in} and $\overline{err}$

$$op := Err_{in} - \overline{err}$$

op is generally positive, as $\overline{err}$ is computed on training data

Definition: average optimism is

$$\omega := E_y[op]$$

we are computing the expected value over the training set outcome

For a reasonable number of loss functions including 0-1 loss and squared error, it can be shown that

$$\omega = \frac{2}{N} \sum_{i=1}^{N} Cov(\hat{y}_i, y_i)$$

where Cov denotes the covariance → Ex 7.4

Note:
- optimism depends on how much y_i affects its own prediction
- the harder we fit the data, the larger the value of $Cov(\hat{y}_i, y_i)$
  → the higher the optimism

As a consequence:

$$E_y[Err_{in}] = E_Y[\overline{err}] + \frac{2}{N} \sum_{i=1}^{N} Cov(\hat{y}_i, y_i)$$

When $\hat{y}_i$ is obtained by a linear fit in the inputs, $Y = f(x) + \varepsilon$

$$\sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = d\,\sigma_\varepsilon^2$$

← effective number of parameters

therefore

$$\rightarrow E_y[Err_{in}] = E_Y[\overline{err}] + 2\frac{d}{N}\sigma_\varepsilon^2 \quad (*)$$

e.g. in linear regression, the number of covariates

- optimism increases with the number of inputs;
- " decreases " " sample size.

Methods we will see:
- C_p, AIC and BIC estimate the prediction error by estimating the training error and the optimism (work when estimates are linear in their parameters)
- cross-validation and bootstrap-based procedure try to estimate directly the expected error

Note:
- in-sample error is generally NOT of interest (we are mainly interested in new data, including new points in X)
- to select the best model / tuning the complexity parameter, we are more interested in the relative difference in errors rather than the absolute one.

Estimates of Err_{in}

Start from (*)

$$E_y[Err_{in}] = E_Y[\overline{err}] + 2\frac{d}{N}\sigma_\varepsilon^2 \quad (*)$$

Write the general form of the in-sample estimate

$$\widehat{Err}_{in} = \overline{err} + \hat{\omega}$$

- when we have linearity and squared errors, from (*)

$$C_p = \overline{err} + 2\frac{d}{N}\hat{\sigma}_\varepsilon^2$$

NB: there are other versions of C_p, different versions may give different numbers, but they all lead to the same model

where:
- $\overline{err}$ is computed through the squared loss;
- d is # of parameters
- $\hat{\sigma}_\varepsilon^2$ is an estimate of the noise variance
  it is computed using the full model, because it has the smallest bias

Similar idea for AIC (Akaike Information Criterion)
- we start again from (*), but we want to be more general (the error is computed through a likelihood approach)

We are using the asymptotic result ($N \to \infty$)

$$-2E[\log P_{\hat\theta}(Y)] \approx -\frac{2}{N}E\Big[\underbrace{\sum_{i=1}^{N}\log P_{\hat\theta}(y_i)}_{\text{loglik}}\Big] + 2\frac{d}{N}$$

- $P_\theta(Y)$ is the family of densities for Y (containing the "true" density)

the maximized log-likelihood $\ell(\hat\theta)$ is the mle maximum likelihood estimate

E.g., in the logistic regression: $AIC = -\frac{2}{N}\text{loglik} + 2\frac{d}{N}$

Gaussian regression: $AIC \propto C_p$ AIC

C_p is a special case of AIC

To find the best model, we choose that with the smallest AIC
- straightforward in the simplest cases; we need more attention in more complex situations. In particular, we need to find a reasonable measure for the model complexity

Note: for regularized/penalized regression

$$AIC(\alpha) = \overline{err}(\alpha) + 2\cdot\frac{d(\alpha)}{N}\hat{\sigma}_\varepsilon^2$$

- usually minimizing AIC is not the best solution to find the value of the tuning parameter $\to$ cross-validation is a better approach

The effective number of parameters

- generalized the concept of number of parameters, in order to extend the previous approaches to more complex situation

Suppose

$y = (y_1, \ldots, y_N)$ outcome

$\hat{y} = (\hat{y}_1, \ldots, \hat{y}_N)$ predictions

Linear methods $\hat{y} = Sy$

where S is a $N \times N$ matrix which:
- depends on X
- independent on Y

Linear regression

$$\hat{y} = \underbrace{X(X^TX)^{-1}X^T}_{S}y$$

Ridge regression

$$\hat{y} = \underbrace{X(X^TX + \lambda I_p)^{-1}X^T}_{S}y$$

Effective number of parameters (or effective degrees of freedom)

$$df(S) = \text{trace}(S)$$

We should replace trace(S) to d to obtain the correct value of the criterion

If $y = f(x) + \varepsilon$, $Var(\varepsilon) = \sigma_\varepsilon^2 \rightarrow \sum_{i=1}^{N} Cov(\hat{y}_i, y_i) = \text{trace}(S)\sigma_\varepsilon^2$

So $df(\hat{y}) = \dfrac{\sum_{i=1}^{N} Cov(\hat{y}_i, y_i)}{\sigma_\varepsilon^2}$ $\rightarrow$ Ex 7.5

The Bayesian approach and BIC

The BIC is an alternative criterion to AIC,
↳ Bayesian Information Criterion

AIC = 2
BIC = log(N)

$$\frac{1}{N}\text{BIC} = -\frac{2}{N}\log \text{lik} + \log N \cdot \frac{d}{N}$$

Despite they are quite similar, AIC and BIC come from completely different ideas. BIC comes from the Bayesian approach to model selection

$$\Pr(\mathcal{M}_m \mid z) \propto \Pr(\mathcal{M}_m)\Pr(z \mid \mathcal{M}_m)$$
$$\propto \Pr(\mathcal{M}_m)\int \Pr(z \mid \mathcal{M}_m, \theta_m)\Pr(\theta_m \mid \mathcal{M}_m)\, d\theta_m$$

To choose between two models, we compare their posterior probabilities

$$\frac{\Pr(\mathcal{M}_m \mid z)}{\Pr(\mathcal{M}_\ell \mid z)} = \frac{\Pr(\mathcal{M}_m)}{\Pr(\mathcal{M}_\ell)} \cdot \frac{\Pr(z \mid \mathcal{M}_m)}{\Pr(z \mid \mathcal{M}_\ell)}$$

Bayes factor

in a large number of case = 1 (give the same prior probability to the two model) ⟹ the choice between the two models is based on the Bayes factor

Approximating

$$\Pr(z \mid \mathcal{M}_m) = \log(z \mid \underset{\text{mle}}{\hat{\theta}_m}, \mathcal{M}_m) - \frac{d_m}{2}\log N + O(1)$$

If the loss function is $-2\log \Pr[z \mid \hat{\theta}, \mathcal{M}_m)$, we obtain BIC

- We may select the model with smallest BIC

 correspond to selecting the model with highest posterior probability

Note that

$$\frac{e^{-\frac{1}{2}\text{BIC}_m}}{\sum_{\ell=1}^{M} e^{-\frac{1}{2}\text{BIC}_\ell}}$$

is the posterior probability of selecting the model m

AIC vs BIC

- no clear preference
- BIC leads to a sparser model
- AIC leads to a model with more predictors
- BIC is consistent ($N \to \infty$, the probability of selecting the true model goes to 1)
- for finite sample size, BIC tends to select a model which is too sparse

Cross-validation

The cross-validation aims to estimate the extra-sample error

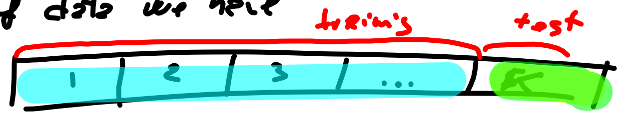
$$Err = E[L(Y, \hat{f}(x))]$$

independent test set

the average test error when $\hat{f}(x)$ is applied to a new sample

If we had enough data → training set
→ test "

↳ in general, we have not, so we mimic this split using the limited amount of data we have



folds

derived using all data but those in fold k

- divide the observations in K folds
- we use, in turn, K-1 folds to train the model (derive $\hat{f}^{-k}(x)$)
- we evaluate the model in the remaining fold

$$CV(\hat{f}) = \sum_{k=1}^{K} \sum_{i=1}^{n_k} L(y_i, \hat{f}^{-k}(x_i))$$

- if $K=2$, two-fold cross-validation
- if $K=N$, leave-one-out cross-validation (LOOCV)
  in this case, each observation is a fold

How do we choose K

- bias-variance trade-off
- smaller the K, larger bias, smaller variance
- larger the K, smaller bias, larger variance
  (the extreme case is LOOCV, where we use N-1 observations for training the model → the training sets are really similar to each other)
- usual choices are $K=5$ or $K=10$

[Fig 7.8]

→ 1-Err vs N
→ the classifier is ok until ≈ $N=100$ (then is flat)

- if $N=200$, $K=5$ → training $N_t = 160$ ✓ $160 > 100$
- if $N=50$, $K=5$ → " $N_t = 40$ ✗ $40 \not> 100$

Notes:

- CV estimates Err and not the Err_{τ}
- if we want to select a tuning parameter via CV

$\hat{f}^{-k}(x, \alpha)$ is the model selected using α and fitted on the observations which do not belong to the k-th fold

$$CV(\hat{f}, \alpha) = \frac{1}{\sum_{k=1}^{K} n_k} \sum_{k=1}^{K} \sum_{i=1}^{n_k} L(y_i, \hat{f}(x_i, \alpha))$$

$$\hat{\alpha} = \arg\min_{\alpha} CV(\hat{f}, \alpha)$$

Generalized cross-validation

- convenient approximation to the LOOCV, for squared loss function

$$\hat{y} = Sy$$

LOOCV

$$\frac{1}{\cancel{\sum_{k=1}^{K} n_k}\; N} \sum_{k=1}^{\cancel{K}\; N} \cancel{\sum_{i=1}^{n_k}} \to \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \hat{f}^{-i}(x_i)\right)^2 = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}} \right)^2$$

where S_{ii} is the ith term on the diagonal of S

The generalized cross-validation (GCV)

$$GCV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{y_i - \hat{f}(x_i)}{1 - \text{trace}(S)/N} \right)^2$$

- computational advantages;
- similarities between AIC and GCV $\to$ Ex 7.7