Penalised Likelihood Part 2: Bias-Variance Tradeoff and Bayesian interpretation

Alex Lewin

Department of Medical Statistics,
London School of Hygiene and Tropical Medicine

Spring 2022

Bias-Variance Trade-Off

So far, have thought about the model and predictions given a particular data set.

Now we start to think about how our fitted model and predictions change when fitted on different data sets.

Suppose the true data-generating model is

$$Y = f(X) + \epsilon$$

with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$

Consider how the estimator $\hat{f}$ varies for different training data sets.

Bias-Variance Trade-Off

MSE at a test data point (x_0, y_0) is $(y_0 - \hat{f}(x_0))^2$

Take Expectation over different training data sets:

$$
\begin{aligned}
EMSE &= E_{Y,X}\left(y_0 - \hat{f}(x_0)\right)^2 \\
&= E_{Y,X}\left(f(x_0) - \hat{f}(x_0) + \epsilon\right)^2 \\
&= \left(E_X(\hat{f}(x_0)) - f(x_0)\right)^2 + E_X\left(\hat{f}(x_0) - E_X(\hat{f}(x_0))\right)^2 + E_{Y|X}(\epsilon^2) \\
&= \left(Bias(\hat{f}(x_0)\right)^2 + Var\left(\hat{f}(x_0)\right) + \sigma^2
\end{aligned}
$$

Bias-Variance Trade-Off

$$EMSE = \left(Bias(\hat{f}(x_0)\right)^2 + Var\left(\hat{f}(x_0)\right) + \sigma^2$$

$Bias(\hat{f}(x_0) = E_X(\hat{f}(x_0)) - f(x_0)$: Bias of fitted model compared to the truth

$Var\left(\hat{f}(x_0)\right)$: Variance of fitted model (amongst different training data sets)

$Var(Y|X) = \sigma^2$: Irreducible error (can't get rid of this for any fitted model)

Bias-Variance Trade-Off

Bias-Variance Trade-Off:

More complex model $\longleftrightarrow$ lower bias, higher variance

Less complex model $\longleftrightarrow$ lower variance, higher bias

Aim to find a balance (trade-off) where both bias and variance are reasonably low.

Penalty parameter λ controls model complexity (like K in KNN, no. variables in subset selection)

Bayesian interpretation

Penalised regression can be seen as a Bayesian method:

$$RSS + Penalty = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

RSS $=$ log Likelihood
Penalty term $=$ log Prior for $\boldsymbol{\beta}$

Ridge/lasso algorithm performs **optimisation** $\longrightarrow$ obtain Maximum A Posteriori (MAP) point estimates (Posterior Mode) for $\boldsymbol{\beta}$

Full Bayesian approach obtains Posterior means, medians, credible intervals on $\boldsymbol{\beta}$, inclusion probabilities etc.

Many possible penalties/priors on β have been proposed

- Ridge generally better for situation where many variables have small effects, possibly correlated
- Lasso better for **sparse** cases: small proportion of variables expected to be important (so good for machine learning/exploratory work)
- Elastic Net penalty is linear combination of ridge and lasso
- Many other shrinkage/regularisation priors on β proposed in Bayesian literature

Choosing between different approaches for variable selection:

$p < n$, small p:
Can use exhaustive subset search (Frequentist using BIC/CV or Bayesian).

$p < n$, moderate p:
Exhaustive search computationally infeasible, so use regularised likelihood or stepwise search.

- Regularisation (penalised likelihood or Bayesian) obtains better regression coefficient estimates.
- Stepwise selection estimates can be biased (but using BIC/CV better than other stepwise approaches).

$p > n$:
Must use regularisation (either penalised likelihood or Bayesian approach).

Further reading:

Lecture notes on ridge and lasso: https://arxiv.org/pdf/1509.09169 Van Wieringen

BayesVerSel R package vignette: https://cran.r-project.org/web/packages/BayesVarSel/index.html

Some of the figures in this presentation may have been taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani