Model Selection via Prior and Posterior over Model Complexity

Motivation

In model fitting, especially with flexible families like polynomials, increasing the number of parameters often reduces training error but can lead to overfitting. Classical cross-validation helps detect overfitting by comparing performance across training and test datasets.

But beyond raw test error, we can **treat model complexity as a random variable** and **assign a probability distribution over it**, allowing us to apply Bayesian reasoning.

Prior and Posterior over Model Degree

We consider a finite set of models indexed by polynomial degree $m = 1, 2, \ldots, m_{\max}$. We place an **exponential prior** over these degrees:

$$\pi(m) = \frac{e^{-\lambda m}}{\sum_{k=1}^{m_{\max}} e^{-\lambda k}}$$

where $\lambda > 0$ favors simpler models.

Given observed test error (quantified by a chi-squared value) for each degree, we define a likelihood-like term for the data under each model:

$$\mathcal{L}(m) \propto \exp\left(-\frac{1}{2\sigma^2}\chi^2(D_B, \theta_A^{(m)})\right)$$

Combining prior and this likelihood, the **posterior over degrees** becomes:

$$P(m \mid D_B) = \frac{\pi(m) \cdot \exp\left(-\frac{1}{2\sigma^2}\chi^2(D_B, \theta_A^{(m)})\right)}{Z}$$

where Z is a normalization constant.

MLE vs MAP

- The **Maximum Likelihood Estimate (MLE)** chooses the degree m minimizing $\chi^2(D_B, \theta_A^{(m)})$.
- The **Maximum A Posteriori (MAP)** estimate chooses the degree m maximizing the posterior $P(m \mid D_B)$. This includes a penalty for model complexity via the prior.

MAP Predictions and Overfitting Control

Using the MAP-selected degree m_{MAP}, we re-fit a new model $\theta_{\text{MAP}}^{(m)}$ using regularized least squares (ridge regression), introducing a prior on parameter magnitudes.

We then compute:

$$\chi^2(D_A, \theta_{\text{MAP}}^{(m)}) \quad \text{(training error with MAP fit)}$$
$$\chi^2(D_B, \theta_{\text{MAP}}^{(m)}) \quad \text{(test error with MAP fit)}$$

These are plotted along with:

$$\chi^2(D_A, \theta_A^{(m)}), \quad \chi^2(D_B, \theta_A^{(m)})$$

to visualize and compare **transferability** of models trained under MLE vs MAP frameworks.

Visualization Summary

The figure `combined_model_fit_and_chi2.png` shows:

- Truth function vs fitted MLE and MAP models
- Chi-squared curves:

$$\chi^2(D_A, \theta_A^{(m)}), \quad \chi^2(D_B, \theta_A^{(m)}), \quad \chi^2(D_B, \theta_{\text{MAP}}^{(m)}), \quad \chi^2(D_A, \theta_{\text{MAP}}^{(m)})$$

- Error bars and variance predictions
- Vertical lines at m_{MLE} and m_{MAP}

Conclusion

The MAP framework incorporates both goodness-of-fit and model simplicity. By examining posterior distributions over model degrees, we gain a probabilistic view of complexity selection and enhance generalization by tempering overfitting tendencies.