INF367A: Probabilistic machine learning

Lecture 8: Model selection and evaluation

Pekka Parviainen

University of Bergen

25.2.2020

Outline

Bayesian model selection

- Marginal likelihood
- BIC

Predictive model selection

- Cross-validation
- AIC

Model checking

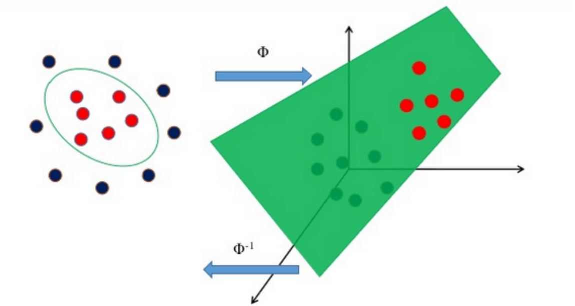
Background

All models are wrong but some are useful

- How to determine which model is most useful?
- How to determine whether a model is useful at all?

Non-linear regression

- Linear models are computationally convenient but as simple models they are unable to capture complex non-linear relations
- Apply a non-linear transformation to data and perform linear regression on transformed data



Source: `https://www.slideshare.net/PeriklisGogas/presentation-machine-learning-56402108`

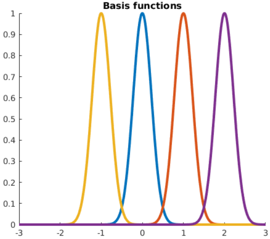
Non-linearities

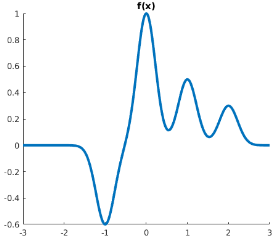
- That is,
$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}),$$
where $\phi(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ is a non-linear transform
- Typically, d' is larger than d
- Computations are done the same way as before but $\mathbf{x}$ is replaced by $\phi(\mathbf{x})$ (and the dimension of $\mathbf{w}$ may change)
- For example, if $\phi(x) = (1, x, x^2)$ then

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ & \cdots & \\ 1 & x_n & x_n^2 \end{pmatrix}$$

Basis functions

- Examples:
  - Polynomials of degree k: $\phi(x) = (1, x, x^2, \ldots, x^k)$
  - Interactions between two features: $\phi((x^{(1)}, x^{(2)})) = (1, x^{(1)}, x^{(2)}, x^{(1)}x^{(2)})$
  - Radial basis functions, e.g., Gaussian: $\phi(\mathbf{x}) = e^{-(\epsilon||\mathbf{x}-\mathbf{c}||)^2}$, where $\mathbf{c}$ is called a center point



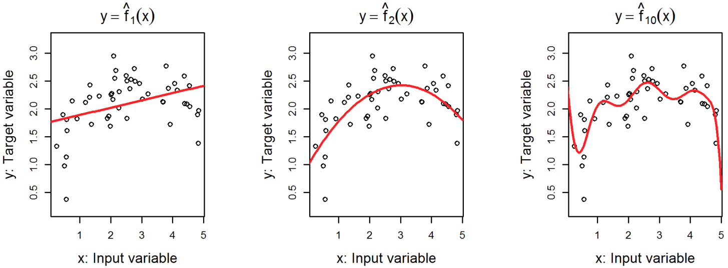


Which set of basis functions should we choose?

- There are lots of possibilities to choose the basis functions
- But which one is good for my particular data?
  - Could use domain knowledge but it is often not available
- $\Rightarrow$ model selection

How to choose a model?

- Choose the most probable/“correct” model
  - Bayesian model selection
  - For example, compare between scientific hypotheses and the hypotheses correspond to different models
- Choose the most predictive model
  - For example, choose the model that best predicts future observations



Bayesian model selection

- Choose the **most probable** model given the observed data
  - Does not evaluate generalisation performance per se
  - Model selection done using the training data
- Here, a model is a family of probability distributions with a fixed set of parameters
  - for example, a family of linear predictors
  - or a family of quadratic predictors
  - . . .

Bayesian model selection

- Consider m models M_i with associated parameters θ_i and associated priors,

$$P(D, \theta_i \,|\, M_i) = P(D \,|\, \theta_i, M_i) P(\theta_i \,|\, M_i) \quad i \in 1, \ldots, m$$

- We can compute *model posterior probabilities*

$$P(M_i \,|\, D) = \frac{P(D \,|\, M_i) P(M_i)}{P(D)}$$

where

$$P(D \,|\, M_i) = \int_{\theta_i} P(D \,|\, \theta_i, M_i) P(\theta_i \,|\, M_i) \mathrm{d}\theta_i$$

and

$$P(D) = \sum_{i=1}^{m} P(D \,|\, M_i) P(M_i)$$

Marginal likelihood

$$P(D \,|\, M_i) = \int_{\theta_i} P(D \,|\, \theta_i, M_i) P(\theta_i \,|\, M_i) \mathrm{d}\theta_i$$

- Also called as *evidence*
- Probability of the data given the model
- The parameters are integrated out

Bayes factor

- Bayesian alternative to classical hypothesis testing (models represent hypotheses)
- The *Bayes factor* is a ratio of marginal likelihoods:

$$BF(M_i, M_j) = \frac{P(D \,|\, M_i)}{P(D \,|\, M_j)}$$

- For compare two, model we compute their posterior ratio:

$$\frac{P(M_i \,|\, D)}{P(M_j \,|\, D)} = \frac{P(D \,|\, M_i)}{P(D \,|\, M_j)} \times \frac{P(M_i)}{P(M_j)}$$

$$\text{Posterior odds} \;=\; \text{Bayes factor} \;\times\; \text{Prior odds}$$

Example

Laplace approximation for marginal likelihood (1/4)

- We want to approximate marginal likelihood

$$P(D \,|\, M_i) = \int P(D \,|\, \theta, M_i) P(\theta \,|\, M_i) \mathrm{d}\theta$$

  when $n \to \infty$

- Laplace approximation for $P(D \,|\, \theta, M_i) P(\theta \,|\, M_i)$:

$$\begin{aligned} E(\theta) &= -\log P(D \,|\, \theta, M_i) P(\theta \,|\, M_i) \\ &= -\sum_{i=1}^{n} \log P(D_i \,|\, \theta, M_i) - \log P(\theta \,|\, M_i) \end{aligned}$$

Laplace approximation for marginal likelihood (2/4)

- We get

$$
\begin{aligned}
\widetilde{E}(\theta) &= E(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T \frac{\partial^2 E(\hat{\theta})}{\partial^2 \theta}(\theta - \hat{\theta}) \\
&= E(\hat{\theta}) + \frac{1}{2}(\theta - \hat{\theta})^T n \cdot F(\hat{\theta})(\theta - \hat{\theta})
\end{aligned}
$$

where

$$
\begin{aligned}
F(\theta) &= \frac{1}{n}\frac{\partial^2 E(\theta)}{\partial^2 \theta} \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\partial^2 \log P(D_i \,|\, \theta, M_i)}{\partial^2 \theta} + \frac{1}{n}\frac{\partial^2 \log P(\theta \,|\, M_i)}{\partial^2 \theta}\Big)
\end{aligned}
$$

- We note that $F(\theta)$ is about an average of second derivatives over all data points

Laplace approximation for marginal likelihood (3/4)

- Let's plugin $\widetilde{E}(\theta)$ back in our formulation:

$$\begin{aligned}
\int P(D \mid \theta, M_i) P(\theta \mid M_i) \mathrm{d}\theta &\approx \int e^{-\widetilde{E}(\theta)} \mathrm{d}\theta \\
&= \int e^{-E(\hat{\theta})} e^{-\frac{1}{2}(\theta-\hat{\theta})^T n \cdot F(\hat{\theta})(\theta-\hat{\theta})} \mathrm{d}\theta \\
&= e^{-E(\hat{\theta})} \int e^{-\frac{1}{2}(\theta-\hat{\theta})^T n \cdot F(\hat{\theta})(\theta-\hat{\theta})} \mathrm{d}\theta \\
&= e^{-E(\hat{\theta})} \left(\frac{2\pi}{n}\right)^{\frac{d}{2}} |F(\hat{\theta})|^{-\frac{1}{2}}
\end{aligned}$$

where the last integration follows from the normalization constant of multivariate Gaussian

Laplace approximation for marginal likelihood (4/4)

- It follows that

$$\begin{aligned}\log P(D \,|\, M_i) &\approx -E(\hat{\theta}) + \frac{d}{2}\log 2\pi - \frac{d}{2}\log n - \frac{1}{2}\log|F(\hat{\theta})| \\ &= \sum_{i=1}^{n} \log P(D_i \,|\, \theta, M_i) + \log P(\theta \,|\, M_i) \\ &+ \frac{d}{2}\log 2\pi - \frac{d}{2}\log n - \frac{1}{2}\log|F(\hat{\theta})|\end{aligned}$$

- Here all other terms expect $\sum_{i=1}^{n} \log P(D_i \,|\, \theta, M_i)$ and $\frac{d}{2}\log n$ converge to a constant when $n \to \infty$

Bayesian Information Criterion (BIC)

- Definition[1]:

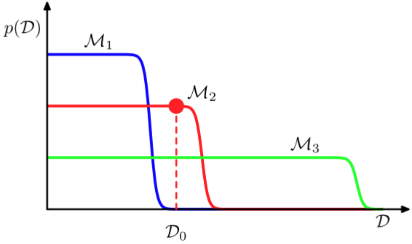
  $$BIC(M) = \log P(D \,|\, \hat{\theta}, M) - \frac{d}{2} \log n$$

  where $\hat{\theta} = \arg\max_{\theta} P(D \,|\, \theta, M) P(\theta \,|\, M)$ are the MAP parameters, d is the number of parameters in model M and n is the sample size
- Large values are good
- Approximates the logarithm of marginal likelihood when $n \to \infty$
- Intuitively, log-likelihood minus a complexity penalty

[1] Sometimes BIC is multiplied by -2

Bayesian model selection and Occam's razor

- When complexity of M increases, $P(D \,|\, \hat{\theta}, M)$ always increases
- On the other hand, $P(D \,|\, M)$ is the highest for the simplest model that can explain the data (= Occam's razor principle)
- Note: Bayesian model selection uses the training set



Source: Bishop, 3.13

Selection models for prediction

- Notation: feature vector $\mathbf{x}$, label y, prediction model $\hat{f}_\theta(\mathbf{x})$ where the parameters θ are estimated from a training data D
- Loss function measure the (lack of) accuracy of prediction
- Squared loss:
$$L(y, \hat{f}_\theta(\mathbf{x})) = (y - \hat{f}_\theta(\mathbf{x}))^2$$
- Loss based on log-likelihood:
$$L(y, \hat{f}_\theta(\mathbf{x})) = -2 \log P(y \,|\, \hat{f}_\theta(\mathbf{x}))$$
(the "-2" makes this match the squared loss for Gaussian models)

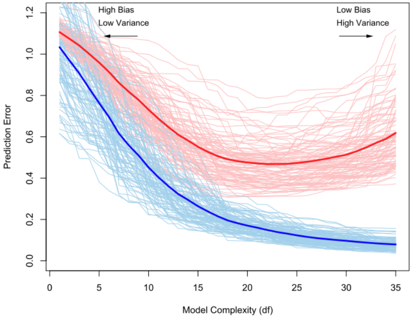
Selection models for prediction

- Test/prediction/generalization error:

$$Err_D = E\left[L(y, \hat{f}_\theta(\mathbf{x}))|D\right]$$

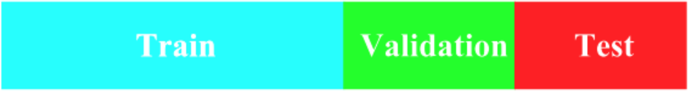
- Training error:

$$\bar{err} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}_\theta(\mathbf{x}_i))$$



Training-validation-testing

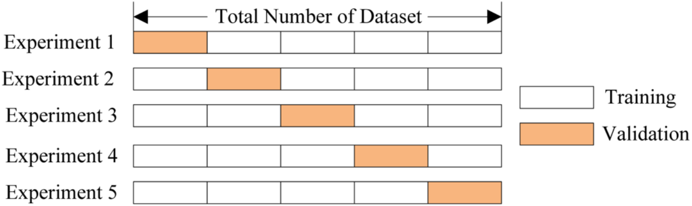
- **Model selection**: estimate the performance of different models in order to choose the best one (use validation data)
- **Model evaluation** (or assessment): estimate the prediction error of the chosen model



- The validation step can be approximated analytically (e.g., AIC) or by efficient sample re-use (e.g., cross-validation)

Example

Cross-validation



- Cross-validation (CV) error:

$$CV(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}^{\kappa(i)}(\mathbf{x}_i))$$

  where $\hat{f}^{\kappa(i)}$ is the predicting model trained without the fold to which observation i belongs to

- CV yields an estimate of the expected prediction error $E[L(y, \hat{f}(\mathbf{x}))]$

Akaike Information Criterion (AIC)

- It can be shown that for **large** n

  $$-2 \cdot E\Big[\log P(\tilde{y}\,|\,\hat{\theta}, \mathbf{x})\Big] = -\frac{2}{n} \log P(y\,|\,\hat{\theta}, X) + 2 \cdot \frac{d}{n}$$

  where $\tilde{y}$ is an unobserved future observation and

  $$\log P(D\,|\,\hat{\theta}) = \sum_{i=1}^{d} \log P(y\,|\,\hat{\theta}, X)$$

  is the log-likelihood
- This leads to

  $$AIC = -\frac{2}{n} \log P(y\,|\,\hat{\theta}, X) + 2 \cdot \frac{d}{n}$$

  (the smaller the better)
- Main point: AIC is one possible analytical approximation for the expected prediction accuracy

BIC vs. AIC vs. cross-validation

- BIC penalizes complexity stronger than AIC
  - BIC tends to select simpler models
- AIC is asymptotically equivalent to leave-one-out cross-validation

Remarks

- Bayesian model selection
  - Asymptotically consistent
  - Suitable when trying to find a “true” model from a set of distinct alternatives
  - Heavy penalty on complexity ⇒ may produce too sparse models for prediction
- Predictive model selection
  - No consistency guarantees
  - No need to assume a true model
  - Less penalty for model complexity ⇒ more complex models that may be more suitable for prediction
- In practice, people use the two ways interchangeably for both goals: prediction and comparing hypotheses

Model checking (1/2)

- We have selected a model. But does it make any sense?
- Whenever possible, perform sanity checks:
  - Use domain knowledge
  - Visualize

Model checking (2/2)

- Residual errors
  - Compute residual errors: $\epsilon_i = y_i - \hat{f}_\theta(\mathbf{x})$
  - Plot the residuals
  - Are your assumptions satisfied? (e.g., if you assume that errors are Gaussians, the histogram of residuals should look like a Gaussian)
- Test statistics
  - Generate data sets (of same size as the original data) from the posterior predictive distribution
  - Compute some descriptive statistics
  - Do the statistics of the generated data sets differ systematically from the original data?

Further readings

- Bishop 3.4, 3.5