Gaussian Processes
Model Selection and Model Assessment

Joanna Ficek

ETH Zurich

23 – 25 October 2019

Tutorial Outline

1. Gaussian Processes

2. Model selection methods

1. Gaussian Processes

2. Model selection methods

Gaussian Processes: recap

Moments of a joint Gaussian:
$\mathbb{E}\left[\mathbf{y}\right] = \mathbf{0}, \mathbb{C}\mathbf{ov}\left[\mathbf{y}\right] = \mathbf{X}\Lambda^{-1}\mathbf{X}^\top + \sigma^2\mathbb{I}_n.$

We can rewrite the joint distribution over $\mathbf{y}$ as follows:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \sim \mathcal{N}\left(\mathbf{y} \middle| \mathbf{0}, \begin{bmatrix} k_{1,1}+\sigma^2 & k_{1,2} & \dots & k_{1,n} \\ k_{2,1} & k_{2,2}+\sigma^2 & \dots & k_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ k_{n,1} & k_{n,2} & \dots & k_{n,n}+\sigma^2 \end{bmatrix}\right)$$

where $k_{i,j} = k(x_i, x_j) := x_i^\top \Lambda^{-1} x_j$ is a kernel function.

- (+) probabilistic approach (estimation and incorporation of uncertainty)
- (+) great flexibility (due to the use of kernels)

Gaussian Processes: kernels

Definition

Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a **kernel** if there exists an $\mathbb{R}$-Hilbert space and a map $\phi : \mathcal{X} \to \mathcal{H}$, such that $\forall \mathbf{x}, \mathbf{x}^{'} \in \mathcal{X}$

$$k(\mathbf{x}, \mathbf{x}^{'}) := \langle \phi(\mathbf{x}), \phi(\mathbf{x}^{'}) \rangle_{\mathcal{H}}$$

Given valid kernels $k_1(\boldsymbol{x}, \boldsymbol{x}^{'})$ and $k_2(\boldsymbol{x}, \boldsymbol{x}^{'})$, the following new kernels are also valid:

1. $k(\boldsymbol{x}, \boldsymbol{x}^{'}) = ck_1(\boldsymbol{x}, \boldsymbol{x}^{'})$, with constant $c > 0$;
2. $k(\boldsymbol{x}, \boldsymbol{x}^{'}) = f(\boldsymbol{x})k_1(\boldsymbol{x}, \boldsymbol{x}^{'})f(\boldsymbol{x}^{'})$, with any function $f(.)$;
3. $k(\boldsymbol{x}, \boldsymbol{x}^{'}) = k_1(\boldsymbol{x}, \boldsymbol{x}^{'}) + k_2(\boldsymbol{x}, \boldsymbol{x}^{'})$;
4. $k(\boldsymbol{x}, \boldsymbol{x}^{'}) = k_1(\boldsymbol{x}, \boldsymbol{x}^{'})k_2(\boldsymbol{x}, \boldsymbol{x}^{'})$;

Gaussian Processes: popular kernels

Kernels defined on $\mathcal{X}$:

- Radial Basis Function (RBF) kernel:

$$\sigma^2 exp(-\frac{||\mathbf{x}-\mathbf{x}^{'}||^2}{2l^2})$$

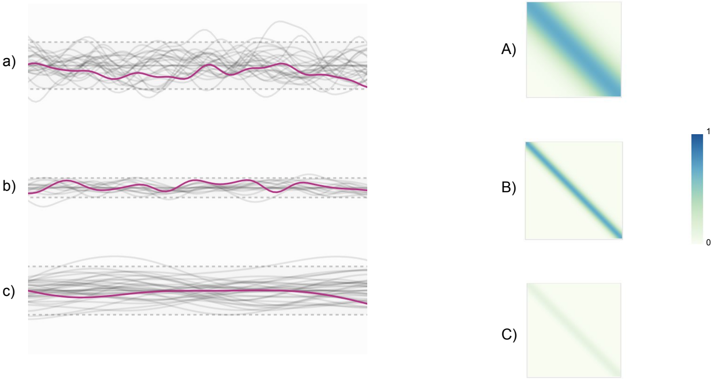
- periodic kernel:

$$\sigma^2 exp(-\frac{2sin^2(\pi|\mathbf{x}-\mathbf{x}^{'}|/p)}{l^2})$$

- linear kernel:

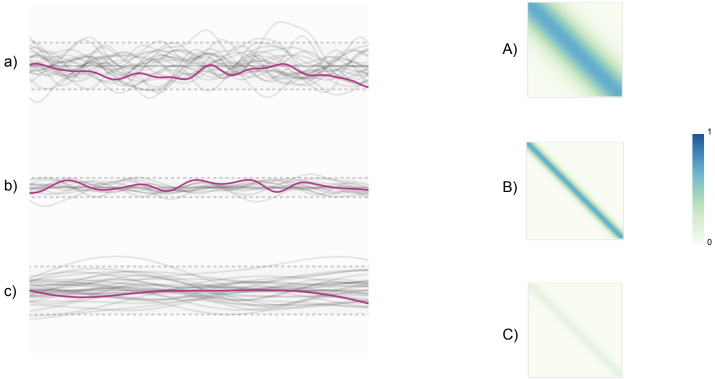
$$\sigma_b^2+\sigma^2(\mathbf{x}-c)(\mathbf{x}^{'}-c)$$

Gaussian Processes: kernels and function shapes



Source: `https://www.jgoertler.com/visual-exploration-gaussian-processes/`.

Gaussian Processes: kernels and function shapes

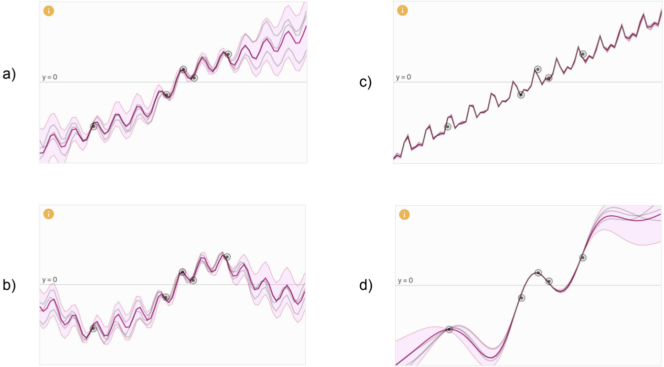


Source: `https://www.jgoertler.com/visual-exploration-gaussian-processes/`.

Figure: Samples from prior distribution obtained using RBF kernel.

1. a:B ($\sigma = 0.8$, $l = 0.5$);
2. c:A ($\sigma = 0.8$, $l = 2$);
3. b:C ($\sigma = 0.33$, $l = 0.5$);

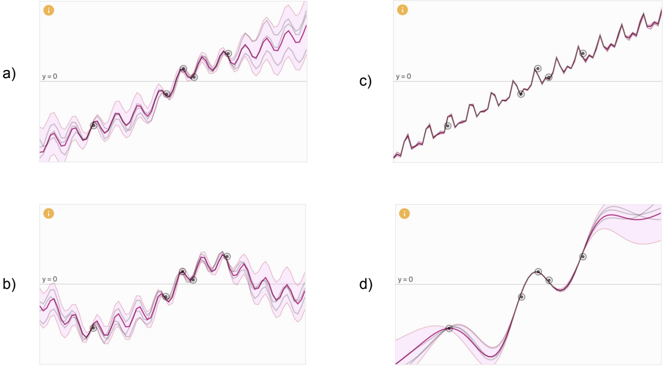
Gaussian Processes: combining kernels



Source: `https://www.jgoertler.com/visual-exploration-gaussian-processes/`.

Figure: Samples from posterior distribution obtained using combinations of different kernels.

Gaussian Processes: combining kernels



Source: https://www.jgoertler.com/visual-exploration-gaussian-processes/.

Figure: Samples from posterior distribution obtained using combinations of different kernels.

a) linear+RBF+periodic; b) RBF+periodic; c) linear+periodic; d) linear+RBF;

1. Gaussian Processes

2. Model selection methods

Model selection

"All models are wrong, but some models are useful"
(Box and Draper 1987)

- Motivation: Estimate generalization error, select best model.
- In a "data-rich" (ideal) situation split data into
  - **train set**: fit models,
  - **validation set**: estimate prediction error for model selection,
  - **test set**: estimate the generalization error;

  What is a "data-rich" situation depends a.o. on signal-to-noise ratio and the complexity of the fitted models.
- Model selection methods can be divided into analytical (AIC, BIC, etc.) and resampling-based (CV, bootstrap).

Notation

- $\{\boldsymbol{x}_i\}$: input data points
- y: target variable
- $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$: training set
- $L(y, y')$: loss function for true target y and predicted target y'
- $\hat{f}_{\mathcal{D}}(\boldsymbol{x}_i)$: prediction model (hypothesis) trained on $\mathcal{D}$

Cross-Validation

1. Split data $\mathcal{D}$ into K sets: $\mathcal{D}_1, \ldots, \mathcal{D}_K$.
2. For $j = 1, \ldots, K$:
   2.1 Train model $\hat{f}_{\mathcal{D} \setminus \mathcal{D}_j}$ on $n\frac{K-1}{K}$ data points, for instance, by minimizing the empirical risk:

$$\hat{f}_{\mathcal{D} \setminus \mathcal{D}_j} := \arg\min_f \frac{1}{|\mathcal{D} \setminus \mathcal{D}_j|} \sum_{i:\boldsymbol{x}_i \notin \mathcal{D}_j} L(y_i, f(\boldsymbol{x}_i)).$$

3. Estimate the prediction error

$$\hat{\mathcal{R}}_{\mathsf{CV}} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}_{\mathcal{D} \setminus \mathcal{D}_z : \boldsymbol{x}_i \notin \mathcal{D}_z}(\boldsymbol{x}_i)).$$

Leave-One-Out Cross-Validation

1. For $i = 1, \ldots, n$:
   1.1 Train model $\hat{f}_{\mathcal{D}\setminus\{(\boldsymbol{x}_i, y_i)\}}$ on $n-1$ data points without $(\boldsymbol{x}_i, y_i)$, for instance, by minimizing the empirical risk:

$$\hat{f}_{\mathcal{D}\setminus\{(\boldsymbol{x}_i, y_i)\}} := \arg\min_f \frac{1}{n-1} \sum_{j:\, j \neq i} L(y_i, f(\boldsymbol{x}_i)).$$

   1.2 Estimate the error of the trained estimator on $(\boldsymbol{x}_i, y_i)$

$$L(y_i, \hat{f}_{\mathcal{D}\setminus\{(\boldsymbol{x}_i, y_i)\}}(\boldsymbol{x}_i)).$$

2. Compute average of the estimated losses

$$\hat{\mathcal{R}}_{\mathsf{LOOCV}} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}_{\mathcal{D}\setminus\{(\boldsymbol{x}_i, y_i)\}}(\boldsymbol{x}_i)).$$

Problem 2: LOOCV for Ridge Regression

- $\boldsymbol{y} \in \mathbb{R}^n$: target values
- $\mathbf{X} \in \mathbb{R}^{d \times n}$: input matrix
- $L(y, y') = (y - y')^2$: loss function

Ridge regression:

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \left[\frac{1}{n} \|\mathbf{X}^T \boldsymbol{w} - \boldsymbol{y}\|^2 + \frac{\mu}{2} \|\boldsymbol{w}\|^2 \right]$$

- $\boldsymbol{y}_{(-i)} \in \mathbb{R}^{n-1}$: all target values except y_i
- $\mathbf{X}_{(-i)} \in \mathbb{R}^{d \times (n-1)}$: input matrix with the i-th column removed

LOOCV:

$$\hat{\mathcal{R}}_{\mathsf{LOOCV}} = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \boldsymbol{x}_i^T \boldsymbol{w}^*_{(-i)}) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i^T \boldsymbol{w}^*_{(-i)} \right)^2$$

Problem 2: LOOCV for Ridge Regression continued

- $\boldsymbol{y} \in \mathbb{R}^n$: target values
- $\mathbf{X} \in \mathbb{R}^{d \times n}$: input matrix

$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \left[\frac{1}{n}\|\mathbf{X}^T\boldsymbol{w} - \boldsymbol{y}\|^2 + \frac{\mu}{2}\|\boldsymbol{w}\|^2\right]$
$\hat{\mathcal{R}}_{\mathsf{LOOCV}} = \frac{1}{n}\sum_{i=1}^n \left(y_i - \boldsymbol{x}_i^T\boldsymbol{w}^*_{(-i)}\right)^2$

Derive:

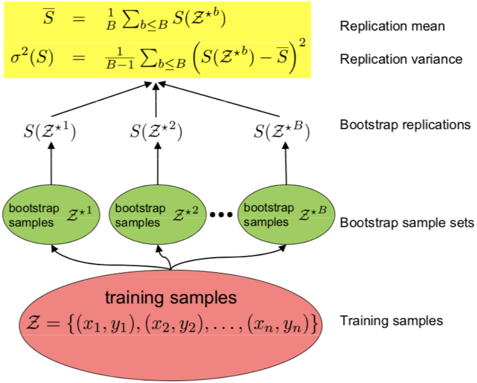
$$\hat{\mathcal{R}}_{\mathsf{LOOCV}} = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - s_i}\right)^2,$$

where

$$\hat{y}_i = \boldsymbol{x}_i^T\mathbf{A}^{-1}\mathbf{X}\boldsymbol{y},$$
$$s_i = \boldsymbol{x}_i^T\mathbf{A}^{-1}\boldsymbol{x}_i,$$
$$\mathbf{A} = \left(\mathbf{X}\mathbf{X}^T + \frac{(n-1)\mu}{2}\mathbf{I}\right).$$

Bootstrap

- Idea: sample n observations with replacement B-times to obtain "additional" data sets and fit the model to each of these data sets



$\overline{S} = \frac{1}{B}\sum_{b\leq B} S(\mathcal{Z}^{\star b})$
$\sigma^2(S) = \frac{1}{B-1}\sum_{b\leq B}\left(S(\mathcal{Z}^{\star b}) - \overline{S}\right)^2$
Replication mean
Replication variance
$S(\mathcal{Z}^{\star 1})$
$S(\mathcal{Z}^{\star 2})$
$S(\mathcal{Z}^{\star B})$
Bootstrap replications
bootstrap samples $\mathcal{Z}^{\star 1}$
bootstrap samples $\mathcal{Z}^{\star 2}$
bootstrap samples $\mathcal{Z}^{\star B}$
Bootstrap sample sets
training samples
$\mathcal{Z} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
Training samples

- Bradley Efron
  `https://www.youtube.com/watch?v=H2tOhMaXWvI`

Bootstrap extensions

Problem: overlap of train and test sets ⇒ too optimistic error estimates.
Solutions (more in the lecture):

- Leave-one-out bootstrap
- 0.632 bootstrap
- $0.632+$ bootstrap

Related topic: stability selection

- Meinshausen and Bühlmann (2009) `https://stat.ethz.ch/~nicolai/stability.pdf`
- Approximate the posterior inclusion probabilities by calculating the frequency of the variable being chosen across bootstrap samples (variable selection).

Model selection: analytical criteria

- Bayesian Information Criterion (**BIC**):

$$-2 \ln p(\mathcal{D}^{(n)}|\theta^{(m)}) + m \ln n$$

  - derived from the Bayesian perspective
  - tends to underfit if small sample size (but asymptotically consistent)

- Minimum Description Length (**MDL**):

$$-\ln p(\mathcal{D}^{(n)}|\theta^{(m)}) - \ln p(\theta^{(m)})$$

  - information theory perspective: "minimize the length of the code to send the message"
  - closely related to BIC, asymptotically equivalent

- Akaike Information Criterion (**AIC**):

$$-2 \ln p(\mathcal{D}^{(n)}|\theta^{(m)}) + 2m$$

  - approximates Kullback-Leibler divergence
  - tends to select complex models if big sample size

Problem 4: Laplace approximation around the mode

Gaussian approximation to a probability density at its mode

- $p(\mathbf{z})$: probability density to approximate
- $\mathbf{z}_0 = \arg\max p(\mathbf{z})$: mode

$$\ln p(\mathbf{z}) \approx \ln p(\mathbf{z}_0) - \frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T \mathbf{A}(\mathbf{z} - \mathbf{z}_0)$$

Note: No first-order term at the mode $\mathbf{z}_0$.

The Laplace approximation around the mode $\mathbf{z}_0$:

$$q(\mathbf{z}) \approx \frac{|\mathbf{A}|^{\frac{1}{2}}}{(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(\mathbf{z} - \mathbf{z}_0)^T \mathbf{A}(\mathbf{z} - \mathbf{z}_0)\right) = \mathcal{N}(\mathbf{z}|\mathbf{z}_0, \mathbf{A}^{-1})$$

Problem 4: normalization constant in Laplace approximation

- $p(\mathbf{z}) = f(\mathbf{z})/Z \propto f(\mathbf{z})$: probability density
- Z: normalization constant

$$\begin{aligned} Z &= \int f(\mathbf{z}) d\mathbf{z} \\ &\approx f(\mathbf{z}_0) \int \exp\left(-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^T \mathbf{A}(\mathbf{z}-\mathbf{z}_0)\right) d\mathbf{z} \\ &= f(\mathbf{z}_0) \frac{(2\pi)^{\frac{n}{2}}}{|\mathbf{A}|^{\frac{1}{2}}} \end{aligned}$$

Evidence appr.: $\mathbf{z} = \theta$, $p(\mathbf{z}) = p(\theta|\mathcal{D})$, $\mathbf{z}_0 = \theta_{\mathsf{MAP}}$, $f(\mathbf{z}) = p(\theta, \mathcal{D})$

$$p(\mathcal{D}) \approx \underbrace{p(\theta_{\mathsf{MAP}}, \mathcal{D})}_{f(\mathbf{z}_0)} \frac{(2\pi)^{\frac{n}{2}}}{|\mathbf{A}|^{\frac{1}{2}}} = \underbrace{p(\mathcal{D}|\theta_{\mathsf{MAP}}) p(\theta_{\mathsf{MAP}})}_{f(\mathbf{z}_0)} \frac{(2\pi)^{\frac{n}{2}}}{|\mathbf{A}|^{\frac{1}{2}}}$$

Problem 4: Bayesian Information Criterion

The Laplace approximation to the log model evidence:

$$\ln p(\mathcal{D}^{(n)}) \approx \ln p(\mathcal{D}^{(n)}|\theta^{(m)}_{\mathsf{MAP}}) + \ln p(\theta^{(m)}_{\mathsf{MAP}}) + \frac{m}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|,$$

where $\mathbf{A} = -\frac{\partial^2}{\partial\theta^{(m)}\partial\theta^{(m)}}\ln p(\mathcal{D}^{(n)}, \theta^{(m)}_{\mathsf{MAP}})$

Assumptions

- $\mathcal{D}^{(n)}$ is iid: $p(\mathcal{D}^{(n)}|\theta^{(m)}_{\mathsf{MAP}}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i|\theta^{(m)}_{\mathsf{MAP}})$
- Hessian $\mathbf{A} \in \mathbb{R}^{m\times m}$ is diagonal

$$[\mathbf{A}]_{jj} = \frac{\partial^2}{\partial\theta_i^2}\ln p(\mathcal{D}^{(n)}, \theta^{(m)}) \sim \sum_{i=1}^{n}\frac{\partial^2}{\partial\theta_j^2}\ln p(\boldsymbol{x}_i|\theta^{(m)}_{\mathsf{MAP}}) \sim nc_j$$

Thus, $\det(\mathbf{A}) \sim \det(Cn\mathbf{I}) = (Cn)^m$ and $\ln|\mathbf{A}| \sim m\ln(Cn) \sim m\ln n$

$\Rightarrow -2\ln p(\mathcal{D}^{(n)}) \approx -2\ln p(\mathcal{D}^{(n)}|\theta^{(m)}) + m\ln n = \mathsf{BIC}(\mathcal{D}^{(n)})$

Jackknife estimator method

Use LOO estimator to estimate the the bias of an estimator

- S: estimated value
- $\hat{S}_n = \hat{S}_n(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$: estimator for S

Leave-One-Out estimator:

$$\hat{S}_{n-1}^{(-i)}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_{i-1}, \boldsymbol{x}_{i+1}, \dots, \boldsymbol{x}_n)$$

Assumption:

$$\mathbb{E}\hat{S}_{n-1}^{(-i)} - S = \sum_{i=1}^{\infty} \frac{a_i}{n^i} = \frac{a_1}{n} + \frac{a_2}{n^2} + \dots$$

Jackknife estimator:

$$\hat{S}_n^{\mathsf{JK}} = \hat{S}_n - (n-1)\left(\frac{1}{n}\sum_{i=1}^{n} \hat{S}_{n-1}^{(-i)} - \hat{S}_n\right)$$

Problem 3: Jackknife estimator

Given:

- $X_1, \ldots, X_n \sim \mathcal{U}[0, \theta]$: samples from uniform distribution
- $X_{(1)}, \ldots, X_{(n)}$: samples sorted in ascending order

Consider the following estimator of the maximum θ:

$$\hat{S}_n = X_{(n)}.$$

1. Prove $\mathbb{E}\hat{S}_n = \frac{n}{n+1}\theta$
2. Write the LOO estimator $\hat{S}_{n-1}^{(-i)}$
3. Compute the Jackknife estimator $\hat{S}_n^{\mathsf{JK}}$
4. Prove that the bias of the Jackknife estimator is less than the bias of the estimator $\hat{S}_n$.

Model selection: general guidelines

- The final prediction error has to be calculated on data **not previously used** neither for model fitting (training) nor for parameter tuning ("model selection");
- Variable selection should be performed within Cross-Validation (see `https://www.youtube.com/watch?v=S06JpVoNaA0`;);
- Cross-Validation is a great framework for model selection, although in each iteration we may underfit the data (due to decreased sample size);
- For unbalanced data use stratified procedures to ensure similar class distribution in training and test sets;
- Do not consider results of iterations as independent, see Bengio and Grandvalet (2004);
- Choice of model selection method depends a.o. on the sample size and the class of the models to be investigated;
- For more guidelines see e.g. Dupuy and Simon (2007) `doi:10.1093/jnci/djk018`;

Literature

- Christopher M. Bishop, *Pattern Recognition and Machine Learning*. Springer Verlag (2006)
- Trevor Hastie, Robert Tibshirani & Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Verlag (2001)
- Kevin P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT Press (2012)
- Carl Edward Rasmussen & Christopher K. I. Williams *Gaussian Processes for Machine Learning*. MIT Press (2006) `http://www.gaussianprocess.org/gpml/`
- `https://www.jgoertler.com/visual-exploration-gaussian-processes/`
- `https://www.cs.toronto.edu/~duvenaud/cookbook/`