SC4/SM4 Data Mining and Machine Learning
Gaussian Processes

Dino Sejdinovic
Department of Statistics
Oxford

Slides and other materials available at:
http://www.stats.ox.ac.uk/~sejdinov/dmml

Gaussian Processes

Parametric vs Nonparametric models

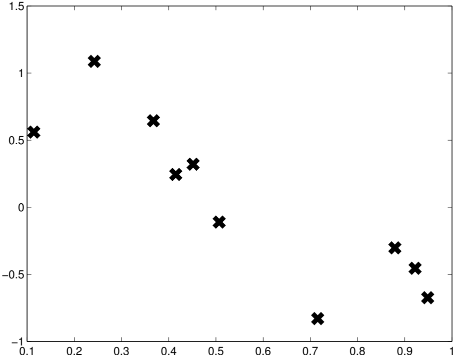
- **Parametric models** have a fixed finite number of parameters, regardless of the dataset size. In the Bayesian setting, given the parameter vector θ, the predictions are independent of the data $\mathcal{D}$.

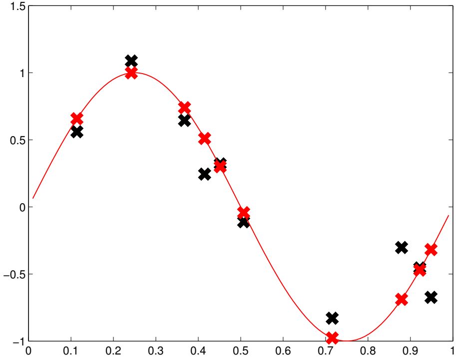
$$p(\tilde{x}, \theta|\mathcal{D}) = p(\theta|\mathcal{D})p(\tilde{x}|\theta)$$

  Parameters can be thought of as a data summary: communication channel flows from data to the predictions through the parameters.

- **Nonparametric models** allow the number of “parameters” to grow with the dataset size. Alternatively, predictions depend on the data (and the hyperparameters).

Regression





- We are given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$.
- Regression: learn the underlying real-valued function $f(x)$.

Different Flavours of Regression

- We can model response y_i as a noisy version of the underlying function f evaluated at input x_i:
$$y_i|f(x_i) \sim \mathcal{N}(f(x_i), \sigma^2)$$
Appropriate loss: $L(y, f(x)) = (y - f(x))^2$
- **Frequentist Parametric** approach: model f as f_θ for some parameter vector θ. Fit θ by ML / ERM with squared loss (linear regression).
- **Frequentist Nonparametric** approach: model f as the unknown parameter taking values in an infinite-dimensional space of functions. Fit f by regularized ML / ERM with squared loss (kernel ridge regression)
- **Bayesian Parametric** approach: model f as f_θ for some parameter vector θ. Put a prior on θ and compute a posterior $p(\theta|\mathcal{D})$ (Bayesian linear regression).
- **Bayesian Nonparametric** approach: treat f as the random variable taking values in an infinite-dimensional space of functions. Put a prior over functions $f \in \mathcal{F}$, and compute a posterior $p(f|\mathcal{D})$ (Gaussian Process regression).

- Just work with the function values at the inputs $\mathbf{f} = (f(x_1), \ldots, f(x_n))^\top$
- What properties of the function can we incorporate?
  - Multivariate normal prior on $\mathbf{f}$:

$$\mathbf{f} \sim \mathcal{N}(0, \mathbf{K})$$

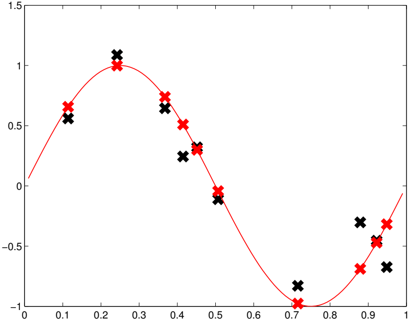
  - Use a kernel function k to define $\mathbf{K}$:

$$\mathbf{K}_{ij} = k(x_i, x_j)$$

  - Expect regression functions to be smooth: If x and x' are close by, then $f(x)$ and $f(x')$ have similar values, i.e. strongly correlated.

$$\begin{pmatrix} f(x) \\ f(x') \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{pmatrix} \right)$$

The prior $p(\mathbf{f})$ encodes our prior knowledge about the function.



- Model:

$$\mathbf{f} \sim \mathcal{N}(0, \mathbf{K})$$

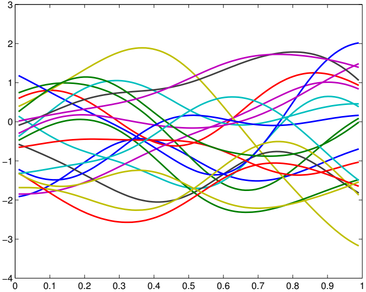
$$y_i | f_i \sim \mathcal{N}(f_i, \sigma^2)$$

Gaussian Processes

- What does a multivariate normal prior mean?
- Imagine $\mathbf{x}$ forms an infinitesimally dense grid of data space. Simulate prior draws

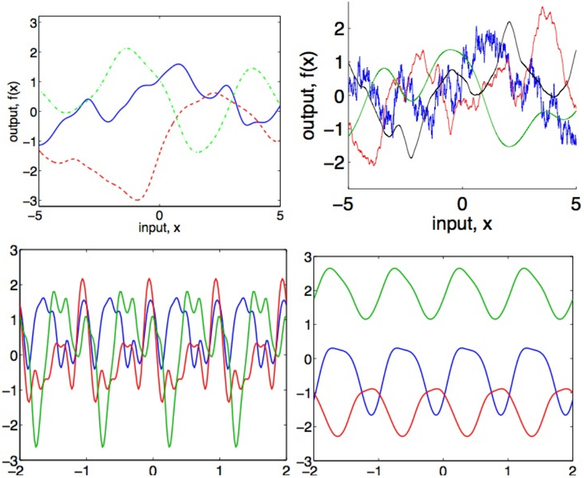
$$\mathbf{f} \sim \mathcal{N}(0, \mathbf{K})$$

  Plot f_i vs x_i for $i = 1, \ldots, n$.
- The corresponding prior over functions is called a **Gaussian Process** (GP): any finite number of evaluations of which follow a Gaussian distribution.



Gaussian Processes

- Different kernels lead to different function characteristics.



Carl Rasmussen. Tutorial on Gaussian Processes at NIPS 2006.

Gaussian Processes

$$\mathbf{f}|\mathbf{x} \sim \mathcal{N}(0, \mathbf{K})$$
$$\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2 I)$$

- Posterior distribution:

$$\mathbf{f}|\mathbf{y} \sim \mathcal{N}(\mathbf{K}(\mathbf{K}+\sigma^2 I)^{-1}\mathbf{y}, \mathbf{K}-\mathbf{K}(\mathbf{K}+\sigma^2 I)^{-1}\mathbf{K})$$

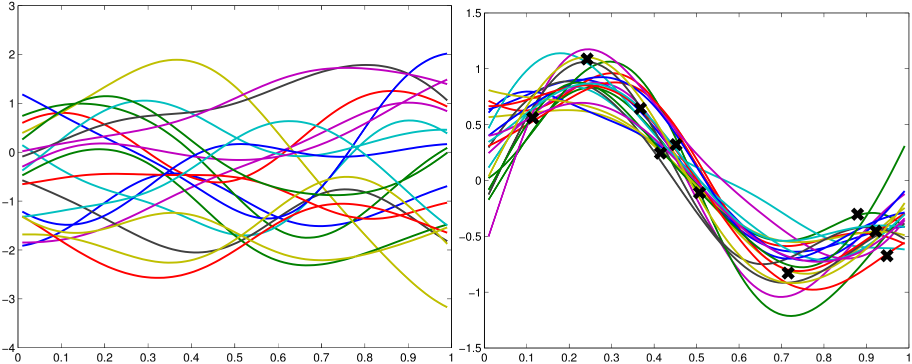
- Posterior predictive distribution: Suppose $\mathbf{x}'$ is a test set. We can extend our model to include the function values $\mathbf{f}'$ at the test set:

$$\begin{pmatrix}\mathbf{f}\\ \mathbf{f}'\end{pmatrix}|\mathbf{x},\mathbf{x}' \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\mathbf{K}_{\mathbf{x}\mathbf{x}} & \mathbf{K}_{\mathbf{x}\mathbf{x}'}\\ \mathbf{K}_{\mathbf{x}'\mathbf{x}} & \mathbf{K}_{\mathbf{x}'\mathbf{x}'}\end{pmatrix}\right)$$
$$\mathbf{y}|\mathbf{f} \sim \mathcal{N}(\mathbf{f}, \sigma^2 I)$$

  where $\mathbf{K}_{\mathbf{x}\mathbf{x}'}$ is matrix with (i,j)-th entry $k(x_i, x'_j)$.
- Some manipulation of multivariate normals gives:

$$\mathbf{f}'|\mathbf{y} \sim \mathcal{N}\left(\mathbf{K}_{\mathbf{x}'\mathbf{x}}(\mathbf{K}_{\mathbf{x}\mathbf{x}}+\sigma^2 I)^{-1}\mathbf{y}, \mathbf{K}_{\mathbf{x}'\mathbf{x}'}-\mathbf{K}_{\mathbf{x}'\mathbf{x}}(\mathbf{K}_{\mathbf{x}\mathbf{x}}+\sigma^2 I)^{-1}\mathbf{K}_{\mathbf{x}\mathbf{x}'}\right)$$

Gaussian Processes



GP regression demo: http://www.tmpl.fi/gp/

- A whirlwind journey through data mining and machine learning techniques:
  - **Unsupervised learning**: PCA, MDS, Isomap, Hierarchical clustering, K-means, spectral clustering, mixture modelling, EM algorithm, collaborative filtering, biclustering.
  - **Supervised learning**: Empirical risk minimisation, logistic regression, support vector machines, kernel methods and Gaussian processes.
  - **Conceptual frameworks**: prediction, performance evaluation, generalisation, overfitting, regularisation, hypothesis spaces, model complexity.
  - **Theory**: statistical learning theory, convex optimisation, Bayesian vs. frequentist learning, parametric vs non-parametric learning.
- **Topics we did not cover**: neural networks and deep learning, generative adversarial training, decision trees and random forests, boosting, semisupervised learning, online learning, reinforcement learning, Bayesian optimisation, probabilistic numerics... we just scratched the surface!
- **Further resources**:
  - Machine Learning Summer Schools, videolectures.net.
  - Conferences: NIPS, ICML, UAI, AISTATS.

Thank You!