Bayesian Kernel Methods

Unit 1: Bayes Rule, Approximate Inference, Hyperparameters

Unit 2: Gaussian Processes, Covariance Function, Kernel

Unit 3: GP: Regression

Unit 4: GP: Classification

Unit 5: Implementation: Laplace Approximation, Low Rank Methods

Unit 6: Implementation: Low Rank Methods, Bayes Committee Machine

Unit 7: Relevance Vector Machine: Priors on Coefficients

Unit 8: Relevance Vector Machine: Efficient Optimization and Extensions

http://mlg.anu.edu.au/~smola/summer2002/

Overview of Unit 2: Gaussian Processes

01: Gaussian Process
02: Inference with Gaussian Processes
03: Example: Finite Dimensional Regression
04: Linear Model
05: Linear Model: Consequences
06: Parametric Family
07: General Covariance Function
08: Example: Diffusion on a Graph
09: Example: Gaussian RBF Kernel
10: Inference: Posterior Distribution
11: MAP Approximation
12: Confidence Intervals
13: Relation to Regularized Risk Functional
14: Coefficient-based Priors

Gaussian Process

Definition

Denote by $t(x)$ a stochastic process parametrized by $x \in \mathcal{X}$ ($\mathcal{X}$ is an arbitrary index set). Then $t(x)$ is a Gaussian process if for any $m \in \mathbb{N}$ and $\{x_1, \ldots, x_m\} \subset \mathcal{X}$, the random variables $(t(x_1), \ldots, t(x_m))$ are normally distributed.

Covariance Function

We denote by $k(x, x')$ the function generating the covariance matrix

$$K := \operatorname{cov}\{t(x_1), \ldots, t(x_m)\} \text{ where } K_{ij} =: k(x_i, x_j).$$

and by μ the mean of the distribution.

Common Assumption: Set $\mu = 0$.

Density at Observations

We observe t at m locations $x_1, \ldots, x_m$. Then $p(\mathbf{t})$ is given by

$$p(\mathbf{t}) = (2\pi)^{-\frac{m}{2}} |K|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{t} - \mu)^\top K^{-1} (\mathbf{t} - \mu)\right)$$

Inference with Gaussian Processes

Goal

After observing $\mathbf{t} := (t(x_1), \ldots, t(x_m))$ we would like to infer the distribution of t at locations $x'_1, \ldots, x'_n$, i.e., we would like to infer about $\mathbf{t}' := (t(x'_1), \ldots, t(x'_n))$.

Conditional Density

We study $p(\mathbf{t}'|\mathbf{t})$. Recall that $p(\mathbf{t}, \mathbf{t}') = p(\mathbf{t}|\mathbf{t}')p(\mathbf{t}')$ and therefore $p(\mathbf{t}|\mathbf{t}')$ can be obtained from $p(\mathbf{t}, \mathbf{t}')$ by **fixing** $\mathbf{t}'$ and **normalizing** by $p(\mathbf{t}') = \int p(\mathbf{t}, \mathbf{t}')d\mathbf{t}$.

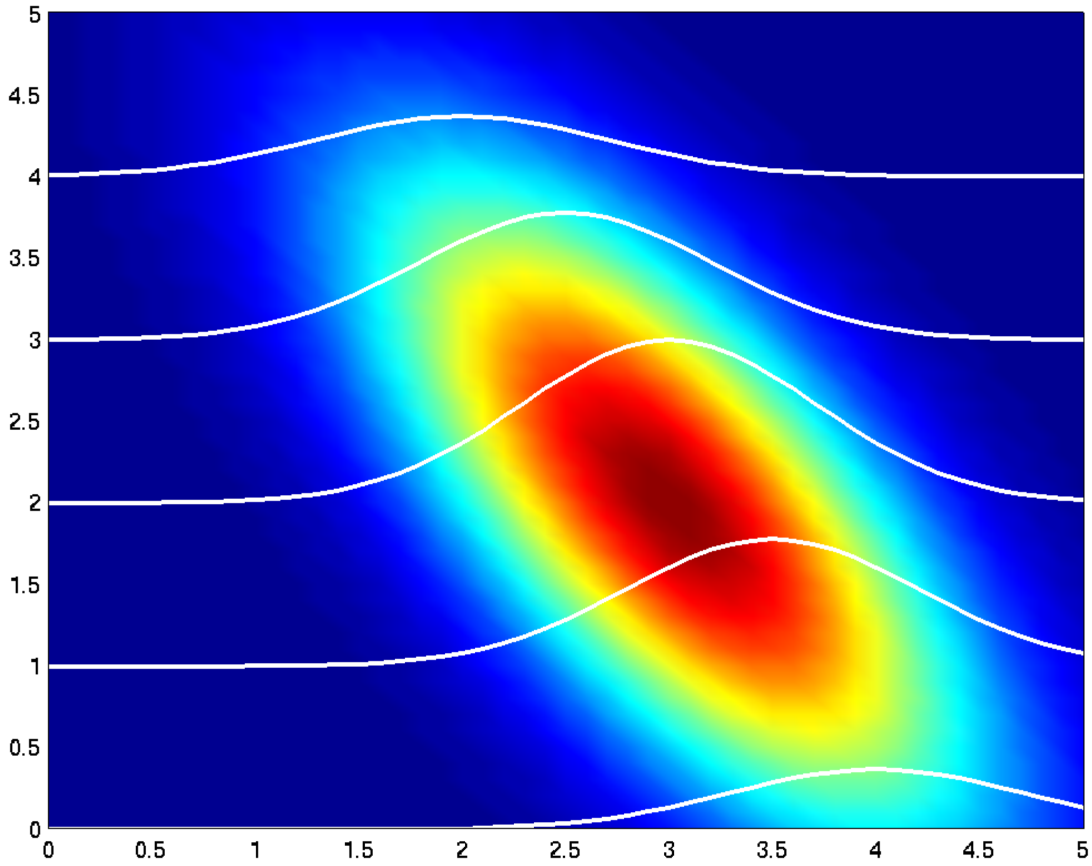
Lazy Trick

For normal distributions we only need to compute **mean** and **covariance** to determine the density completely (including normalization factors).

Recipe: collect all terms from $p(\mathbf{t}, \mathbf{t}')$ dependent on $\mathbf{t}'$ and ignore the rest.

$$p(\mathbf{t}, \mathbf{t}') \propto \exp\left(-\frac{1}{2}\left(\begin{bmatrix}\mathbf{t}\\ \mathbf{t}'\end{bmatrix} - \begin{bmatrix}\mu\\ \mu'\end{bmatrix}\right)^\top \begin{bmatrix}K_{\mathbf{tt}} & K_{\mathbf{tt}'}\\ K_{\mathbf{t}'\mathbf{t}} & K_{\mathbf{t}'\mathbf{t}'}\end{bmatrix}^{-1} \left(\begin{bmatrix}\mathbf{t}\\ \mathbf{t}'\end{bmatrix} - \begin{bmatrix}\mu\\ \mu'\end{bmatrix}\right)\right)$$

Example: Regression without Noise



Example: Regression without Noise

Inverting the Covariance Matrix

$$\begin{bmatrix} K_{\mathbf{tt}} & K_{\mathbf{tt}'} \\ K_{\mathbf{tt}'}^{\top} & K_{\mathbf{t}'\mathbf{t}'} \end{bmatrix}^{-1} = \begin{bmatrix} {K_{\mathbf{tt}}}^{-1} - \left({K_{\mathbf{tt}}}^{-1}K_{\mathbf{tt}'}^{\top}\right)^{\top}\chi^{-1}\left({K_{\mathbf{tt}}}^{-1}K_{\mathbf{tt}'}^{\top}\right) & -\left({K_{\mathbf{tt}}}^{-1}K_{\mathbf{tt}'}^{\top}\right)\chi^{-1} \\ -\chi^{-1}\left({K_{\mathbf{tt}}}^{-1}K_{\mathbf{tt}'}^{\top}\right)^{\top} & \chi^{-1} \end{bmatrix}$$

where $\chi = K_{\mathbf{t}'\mathbf{t}'} - K_{\mathbf{tt}'}^{\top}{K_{\mathbf{tt}}}^{-1}K_{\mathbf{tt}'}$ (Schur complement).

Reduced Covariance

From the inverse of the covariance matrix we obtain that the only quadratic part in $\mathbf{t}'$ is given by χ. Thus the **variance in $\mathbf{t}'$ is y reduced** from $K_{\mathbf{t}'\mathbf{t}'}$ to $K_{\mathbf{t}'\mathbf{t}'} - K_{\mathbf{tt}'}^{\top}{K_{\mathbf{tt}}}^{-1}K_{\mathbf{tt}'}$ by observing $\mathbf{t}$.

Predictive Mean

Instead of μ' the mean is shifted to $\mu' + K_{\mathbf{tt}'}^{\top}K_{\mathbf{tt}}^{-1}(\mathbf{t} - \mu)$.

Linear Model

Covariance Function

Assume that $\mathrm{Cov}(t(x), t(x')) = \langle x, x' \rangle$ with $x \in \mathbb{R}^n$, i.e., that we have an n-dimensional Normal distribution, where the covariance between observations is a bilinear function of x and x'.

Density

$$p(\mathbf{t}) = (2\pi)^{-\frac{n}{2}} \left(\det X^\top X\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{t}-\mu)(XX^\top)^*(\mathbf{t}-\mu)\right)$$

where $X = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ and $(XX^\top)^*$ is the pseudoinverse of $XX^\top$.

Parameter Transformation

By letting $\mathbf{t} = X\alpha + \mu$ (this is admissible since $p(\mathbf{t})$ only defined a density on an n-dimensional subspace) we see that this is equivalent to

$$p(\alpha) = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\|\alpha\|^2\right) \text{ where } \mathbf{t} = X\alpha + \mu.$$

see e.g., Box and Tiao, 1973.

Linear Model, Part II

Prediction

Since $\mathbf{t} = X\alpha + \mu$, already after observing $m = n$ instances $\{x_1, \ldots, x_n\} \subset \mathbb{R}^n$ we can determine α completely.

Reason: X spans only an n-dimensional subspace.

Advantage

We only need n observations.

Problem 1

The model breaks if $\mathbf{t} \neq X\alpha + \mu$ for all $\alpha \in \mathbb{R}^n$. We need to modify our statistical model.

Problem 2

We may have an overly simple model, so we cannot learn beyond a certain point.

Parametric Family

Extension

Instead of $k(x, x') = \sum_{i=1}^{m} x_i x'_i$ we assume the covariance function

$$k(x, x') = \sum_{i=1}^{N} \phi_i(x)\phi_i(x').$$

where $\phi_i(x)$ are the features.

Reparametrization

As in the linear case reparametrize $\mathbf{t} = \Phi\alpha$, where $\Phi_{ij} = \phi_i(x_j)$. Therefore we have **two equivalent parametrizations** of the prior on $\mathbf{t}$ (assuming $m \geq N$):

$$\begin{aligned} p(\alpha) &= (2\pi)^{-\frac{N}{2}} \exp\left(-\frac{1}{2}\|\alpha\|^2\right) \text{ and } \mathbf{t} = \Phi\alpha + \mu. \\ p(\mathbf{t}) &= (2\pi)^{-\frac{N}{2}} \left(\det\Phi^\top\Phi\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{t}-\mu)^\top(\Phi\Phi^\top)^*(\mathbf{t}-\mu)\right). \end{aligned}$$

See e.g., Fahrmeir and Tutz, 1994.

General Covariance Function

Idea

In general, we may not know how many dimensions the function space, or, in other words, the space of observations really has, hence use generic kernel k without further assumptions on the dimensionality of the set of functions $k(x_i, \cdot)$.

Examples

$$
\begin{aligned}
k(x, x') &= \exp\left(-\frac{1}{2\sigma\|x - x'\|}\right) \text{ Laplacian Kernel} \\
k(x, x') &= \exp\left(-\frac{1}{2\sigma^2\|x - x'\|^2}\right) \text{ Gaussian RBF Kernel} \\
k(x, x') &= \left(\langle x, x' \rangle + c\rangle\right)^d \text{ with } c \geq 0,\ d \in \mathbb{N} \text{ Polynomial Kernel} \\
k(x, x') &= B_{2n+1}(x - x') \text{ Spline kernel} \\
k(x, x') &= \mathbf{E}_c[p(x|c)p(x'|c)] \text{ Conditional Expectation Kernel}
\end{aligned}
$$

All these kernels correspond to a Gaussian process ... (see Williams 1998, Schölkopf and Smola 2002, Wahba 1990,

Diffusion Process

Basic Idea

We have an initial density $p(x, 0)$ of particles, heat, etc., which becomes more spread out over time due a diffusion process. Goal: estimate $p(x, t)$, based on $p(x, 0)$.

Diffusion in $\mathbb{R}$

The change in density is proportional to the second derivative of $p(x, t)$

$$\partial_t p(x, t) = \sigma \partial_x^2 p(x, t)$$

We want to find solutions of the homogeneous PDE.

Extension

More generally we assume a differential equation $\partial_t p(x, t) = D p(x, t)$ where D is a differential operator whose characteristic polynomial of D satisfies $D(\xi) = D(-\xi)$.

Example

Standard diffusion process: $Dp(x, t) = \sigma \Delta p(x, t)$ and correspondingly $D(\xi) = \xi^2$
Likewise $D = 1 + \partial_x^2 + c\partial_x^4$ and $D(\xi) = 1 + \xi^2 + c\xi^4$.

Diffusion Process, part II

Symbolic Solution

We may write $p(x,t) = \exp(Dt)p(x,0)$, which leads to

$$\partial_t p(x,t) = \partial_t \exp(Dt)p(x,0) = D\exp(Dt)p(x,0) = Dp(x,t)$$

Explicit Solution We use the Fourier representation of D and p to obtain

$$\partial_t \mathfrak{F}[p](\omega,t) = D(i\omega)\mathfrak{F}[p](\omega,t)$$

The homogeneous solution $p(x,t)$ is therefore given by

$$p(x,t) = \left(\mathfrak{F}^{-1}[\exp(tD(i\omega))]\right) \circ p(x,0)$$

Example: Diffusion in $\mathbb{R}$

We have $D = \partial_x^2$ and consequently $D(i\omega) = -\omega^2$. This leads to

$$\left(\mathfrak{F}^{-1}[\exp(tD(i\omega))]\right) = \left(\mathfrak{F}^{-1}[\exp(-t\omega^2)]\right) = \frac{1}{\sqrt{4\pi t}}\exp\left(-\frac{x^2}{4t}\right)$$

See e.g. Kondor 2002, Haken, 1976

Diffusion Process, part III

Joint Covariance Function

The function $G_t(x) := \left(\mathfrak{F}^{-1}[\exp(tD(i\omega))]\right)(x)$ gives the density of observing a particle at location x, if we started with all the probability mass located at $x = 0$ at time $t = 0$. Hence, the joint probability of observing particles at x, x' is given by

$$p(x, x'|t, x_{\text{start}} = 0) = G_t(x)G_t(x')$$

Uniform Initialization: assuming that at time $t = 0$ the density is uniform, we have

$$\begin{aligned} p(x, x') &= \int G_t(x - \tau)G_t(x' - \tau)d\tau \\ &= (G_t \circ G_t)(x - x')(\text{ Symmetry in } G_t) \\ &= \left(\mathfrak{F}^{-1}[\exp(2tD(i\omega))]\right)(x - x') = G_{2t}(x - x')(\text{Fourier-Plancherel}). \end{aligned}$$

Simplifying Conclusion

The logarithm of the Fourier transform of a translation invariant kernel corresponds to the differential operator of the generating diffusion process.

Example: Diffusion on a Graph

Connectivity Matrix

Assume an undirected graph with m nodes, then we can represent it by a matrix $C \in \mathbb{R}^{m \times m}$ where $C_{ij} = 1$ if i, j are connected and $C_{ij} = 0$ otherwise.

Next denote by $L := G - \operatorname{diag}(\mathbf{l})$ where $l_i := \sum_j G_{ij}$ the Laplacian of the graph G.

Random Walk on a Graph

Assume that we have a probability distribution on a graph, given by $p \in \mathbb{R}^m$, where $\|p\|_1 = 1$. During time Δt a fraction of $\sigma \cdot \Delta t$ will move from node i to each of the adjacent connected nodes j. This implies that

$$p_i \leftarrow p_i - \sigma \Delta t p_i \sum_j C_{ji} + \sigma \Delta t \sum_j C_{ij} p_j = p_i + \sigma \Delta t [Lp]_i$$

Limiting Case (Kondor, 2002)

After n steps the density p becomes $(1 + \sigma \Delta t L)^n p$. If we now set $\Delta t = \frac{t}{n}$ and let $n \to \infty$, we obtain

$$p = \lim_{n \to \infty} \left(1 + \frac{\sigma t}{n} L\right)^n = \exp(t \sigma L).$$

Inference: Posterior Distribution

Recall: Bayes Rule

Given X we want to infer $p(f|X,Y)$. With the usual assumptions (iid data, prior independent of X) this leads to

$$p(f|X,Y) \propto p(Y|f,X)p(f) = \prod_{i=1}^{m} p(y_i|f(x_i),x_i)p(f)$$

GP Assumption

The function values $f(x_i)$ are distributed according to a Gaussian process. The connection to the observations y_i is take care by the noise model $p(y_i|f(x_i),x_i)$. This leads to the following log-posterior

$$-\log p(f|X,Y) = \sum_{i=1}^{m} -\log p(y_i|x_i,f(x_i)) + \frac{1}{2}\log\det K + \frac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{f} + c$$

Inference

Inference by computing e.g., $y = \mathbf{E}_{p(f|X,Y)}[f(x)]$ or $\sigma^2 = \mathbf{E}_{p(f|X,Y)}[(f(x)-y)^2]$.

MAP Approximation

Problem

Computing integrals is expensive, in particular in high-dimensional spaces.

MAP Solution

Approximate $\mathbf{E}_{p(f|X,Y)}[f] \approx \mathrm{argmax}_{\mathbf{f}} p(f|X,Y)$. In the present case this means that we solve

$$\underset{\mathbf{f}}{\mathrm{argmin}} \sum_{i=1}^{m} -\log p(y_i|x_i, f(x_i)) + \frac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{b} + c$$

Reparametrization

Set $y = K\alpha$. This leads to the optimization problem

$$\underset{\alpha}{\mathrm{argmin}} \sum_{i=1}^{m} -\log p([K\alpha]_i|x_i, f(x_i)) + \frac{1}{2}\alpha^\top K\alpha + c$$

Prediction

Once we obtained α for X, Y, we may predict $f(x')$ as $\sum_{i=1}^{m} k(x_i, x')\alpha_i$.

This assumes that $\alpha' = 0$ is a good estimate.

Latent Variables

Problem

$\alpha' = 0$ is often not such a good estimate.

This is especially the case if $-\log p(y|x, f(x))$ does not have a minimum (e.g., loss for classification).

Better Solution

Find f such that the expected log-posterior (with the expectations taken over $y'_1, \ldots y'_{m'}$, and adjusted by themselves to minimize the log-posterior) is minimized.

$$\operatorname*{argmin}_{\mathbf{f}, p(\mathbf{y}')} \sum_{i=1}^{m} -\log p(y_i|x_i, f(x_i)) - \mathbf{E}_{y'_1, \ldots, y'_{m'}} \sum_{i=1}^{m'} \log p(y'_i|x'_i, f(x'_i)) + \frac{1}{2}\mathbf{f}^\top K^{-1}\mathbf{f} + c$$

where K is the covariance matrix over X, X' and likewise $\mathbf{f} \in \mathbb{R}^{m+m'}$.

Algorithm (EM, compare to SVM Transduction)

1) For fixed $p(\mathbf{y}')$ find optimal $\mathbf{f}$ (Maximization).

2) For fixed $\mathbf{f}$, find optimal $p(\mathbf{y}')$ (Expectation).

Confidence Intervals

Normal Distribution

If the predictive distribution is a normal distribution, we only need to compute the variance of $y'_1, \ldots y'_{m'}$ to obtain error bars on the prediction (see the reasoning before). Moreover, the MAP approximation is exact.

y'_i have Finite Cardinality

For instance, if we want to predict class labels, we can simply evaluate $p(y = 1|f, x)$ and $p(y = -1|f, x)$ to obtain information about the confidence of the estimate.

General Case: Approximations

Often $p(y|f, x)$ will be none of the above, and, in particular, we will not be able to compute the integrals explicitly, so we have to approximate:

- Quadratic approximation: compute Taylor expansion of $p(f|X, Y)$ at f_{MAP} and use the latter to approximate $p(f|X, Y)$ by a normal distribution.
- Monte Carlo method: sample from $p(f|X, Y)$ (not topic of the lectures here).

Coefficient-based Priors

Factorizing Priors

Analogously to a factorizing assumption on the observations we may also assume

$$p(f) = \prod_{i=1}^{m} p(\alpha_i) \text{ where } f = \sum_{i=1}^{m} \alpha_i f_i$$

Motivation

The basis functions f_i correspond to independent "factors" causing the observations, e.g., neurons firing independently but rarely, image elements occurring, etc.

Example: Laplace Prior

Sparse codes are often represented by $p(\alpha_i) = \frac{1}{2}\exp(-|\alpha_i|)$. Often one uses a distribution which is even more peaked at 0 to obtain a posterior with higher sparsity (e.g., the adjoint Bessel function from before).

Example: Normal Prior

Priors such as $p(\alpha_i) = (2\pi)^{-\frac{1}{2}} \exp(-\frac{1}{2}\alpha_i^2)$ lead to Gaussian processes.