AE8803 Assignment #2 Scratch Work

Oh!! Gaussian Process building

So... basic formula:

$\vec{y} = f(\vec{x}) + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 I)$

$\Rightarrow p(y|f,\sigma^2) \sim N(f(x), \sigma^2)$

then, $p(f|y, X, \sigma^2) =$ posterior

, y = measurements (noisy)
taken @ x
$F(x)$ = true function of x
ε = noise

say $F(x) \sim N(0, \Sigma)$
↳ prior

Standard Linear Model

$y = f(x) + \varepsilon = X^T w + \varepsilon$, $\varepsilon \sim N(0, \sigma_n^2 I)$

$w \sim N(0, \Sigma_0)$

$x = \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \end{pmatrix} \rightarrow X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & & x_{1n} \\ x_{12} & x_{22} & & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{1D} & x_{2D} & & x_{nD} \end{pmatrix}$ (n × D)

add mean & variance of Gaussians
(ε only varying parameter, mean = 0)

$p(y|X, w, \sigma^2) = N(X^T w, \sigma_n^2 I)$
↳ likelihood.

prior: $p(w) = N(0, \Sigma_w)$

posterior: $p(w|X, y, \sigma^2) = \frac{p(y|X, w, \sigma^2)\, p(w)}{p(y|X)} = \frac{p(y, w|X)}{p(y|X)}$

$p(w|X, y, \sigma^2) = \frac{p(w, y|X, \sigma^2)}{p(y|X, \sigma^2)}$ → conditional probability identity

$= \frac{p(y|X, w, \sigma^2)\, p(w)}{p(y|X, \sigma^2)}$ → Bayes' Theorem

$= \frac{p(y|X, w, \sigma^2)\, p(w)}{\int p(y|X, w, \sigma^2)\, p(w)\, dw}$ → marginal

→

$\int p(y|X,w,\sigma^2)\, p(w)\, dw = p(y|X,\sigma^2)$

now, since we are investigating $p(w|y,X,\sigma^2)$, then $p(y|X,\sigma^2)$ will be a constant that normalize $p(w|y,X,\sigma^2)$

thus, $p(w|y,X,\sigma^2) \propto$ likelihood × prior

$= p(y|X,w,\sigma^2)\, p(w)$

$= N(X^T w, \sigma^2)\, N(\mu_0, \Sigma_0)$

$= ~~N(\ldots, \sigma^2)~~$

$\boxed{= N(\mu_w, \Sigma_w)}$

$\mu_w = \Sigma_w \left(\frac{1}{\sigma^2} X y + \Sigma_0^{-1} \mu_0\right)$

$\Sigma_w = \left(\frac{1}{\sigma^2} X X^T + \Sigma_0^{-1}\right)^{-1}$

Now, the posterior predictive pdf is:

~~the marginalization of $p(f^*, w|X^*, X, y)$~~

$p(f^*|X_*, X, y) = \int p(f^*|w, X^*)\, p(w|X,y)\, dw$ (marginal)

$p(f^*|w, X^*) = N(X^{*T} w, 0)$ → by change of ~~variables~~, get mean = 0

$p(w|X,y)$ = posterior = $N(\mu_w, \Sigma_w)$

↳ similar to sum of two Gaussians: $Z = X + Y$

$\Rightarrow Z \sim N(\mu_x + \mu_y, \Sigma_x + \Sigma_y)$

but b/c

$F = X^T w \Rightarrow$

$F \sim N(X^T \mu_w, X \Sigma_w X^T)$

$\Rightarrow p(f^*|X_*, X, y) \sim N(X^{*T} \mu_w, X^{*T} \Sigma_w X^*)$

Now, if we let $\phi(x)$ replace x as a basis function, to make the model more general, then we have

$$y = \phi(x)^T w + \varepsilon \Rightarrow \vec{y} = \Phi(x)^T w + \varepsilon \quad , x \in \mathbb{R}^D$$

$$\Phi(x) = [\phi(x_1) \ \phi(x_2) \dots \phi(x_N)] \in \mathbb{R}^{N \times n}$$

$$\phi(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_N(x) \end{bmatrix}$$ ← each entry is a function of vector x (input vector)

$$x_i = \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iD} \end{bmatrix}, x_i \in \mathbb{R}^D$$

now, N-dimensional design matrix.

$$\Rightarrow \boxed{w \in \mathbb{R}^{N \times 1}}$$

Same results really...

let $p(w) = N(\mu_0, \Sigma_0)$ (prior)

$p(y|X, w, \sigma^2) = N(\Phi(x)^T w, \sigma_n^2 I)$ (likelihood)

$p(w|y, X, \sigma^2) = N(\mu_w, \Sigma_w)$ (posterior)

$$\mu_w = \Sigma_w \left(\frac{1}{\sigma^2} \Phi(x) y + \Sigma_0^{-1} \mu_0 \right)$$

$$\Sigma_w = \left(\frac{1}{\sigma^2} \Phi(x) \Phi(x)^T + \Sigma_0^{-1} \right)^{-1}$$

then

$$\Rightarrow p(y_{new} | x_{new}, X, y) = N\left(\Phi(x)_{new}^T \mu_w, \ \Phi(x_{new})^T \Sigma_w \Phi(x_{new}) \right)$$

$n_1 \times N \quad N \times N \quad N \times n_2 \quad = n_1 \times n_2$ ✓ (posterior predictive)

If we let $K(X_1, X_2) = \Phi(X_1) \Sigma_w \Phi(X_2)$ then can re-write as:

$$p(y_{new} | x_{new}, X, y) = N\left(K(X_{new}, X) (K(XX) + \sigma_n^2 I)^{-1} y, \right.$$

$$\left. K(x_{new}, x_{new}) - K(x_{new}, X)(K + \sigma_n^2 I)^{-1} K(X, x_{new}) \right)$$

Oh. so now we're actually looking @ Gaussian processes
defined as a collection of random var.

let $f(x) \sim GP(m(x), k(x,x'))$ be the set of random variables representing a Gaussian Process.

note: $E[f(x)] = m(x)$

$\text{Var}[k] = E[(f(x)-m(x))(f(x')-m(x'))^T] = k(x,x')$

as a simple example,

$f(x) = \phi(x)^T w$ w/ $w \sim N(0, \Sigma_p)$

$\Rightarrow E[f(x)] = E[\phi(x)^T w] = \phi(x)^T E[w] = 0$

$E[(f(x)-m(x))(f(x')-m(x'))^T]$

$= E[f(x) f(x')^T] = E[\phi(x)^T w w^T \phi(x')] = \phi(x)^T \Sigma_p \phi(x') = \boxed{k(x,x')}$

if y is noisy training data, then $y = f(x) + \varepsilon$, $\varepsilon \sim N(0, \sigma_n^2 I)$

$\Rightarrow y \sim N(0, k(x,x) + \sigma_n^2 I)$

and for a joint distribution w/ new test points X^*;

$$\begin{bmatrix} y \\ f^* \end{bmatrix} \sim N\left(0, \begin{bmatrix} k(x,x)+\sigma_n^2 I & k(x, x_*) \\ k(x_*, x) & k(x_*, x_*) \end{bmatrix}\right)$$

$\rightarrow f_* | y, X, X_* \sim N(\bar{f}_*, \text{cov}(f_*))$ where (per Gaussian conditional derivation) →

$\bar{f}_* = k(x_*, x)[k(x,x) + \sigma_n^2 I]^{-1} y$

$\text{cov}(f_*) = k(x_*, x_*) - k(x_*, x)[k(x,x) + \sigma_n^2 I]^{-1} k(x, x_*)$

if we want to know y_*, then add $\sigma_n^2 I$ to f_* covariance

$$\Rightarrow \boxed{y_* | y, X, X_* \sim N(\bar{f}_*, \text{cov}(f_*) + \sigma_n^2 I)}$$

$k(x, x')$ is the "kernel"

an example of a kernel is the Squared Exponential kernel:

$$k(x_i, x_j) = \alpha \exp\left(-\frac{1}{2l^2} \|x_i - x_j\|_2^2\right)$$

α: amplitude, l: length-scale

$\exp$: sum of infinite # of exponentials...

$\|x_i - x_j\|_2^2$: i.e. distance between two points → "stationary"; also "isotropic" → absolute distance between points to the norm of distance

linear:

$$k(x, x') = \alpha x^T x' \qquad \alpha > 0$$

polynomial

$$k(x, x') = \alpha (1 + x^T x')^\gamma \qquad , \alpha, \gamma > 0$$

cosine

$$k(x, x') = \alpha^2 \cos\left(\frac{2\pi}{l^2} \|x - x'\|_2^2\right) \qquad , \alpha, l > 0$$

Kernels can be combined to capture different aspects:

product $k(x,x') = k_A(x,x')\, k_B(x,x')$

scale $k(x,x') = g(x)\, k(x,x')\, g(x')$, arbitrary $g(x)$

sum $k(x,x') = k_A(x,x') + k_B(x,x')$

For our case, we have kernel that is like:

periodic ← should we have multiple?...

polynomial ← squared?

noise ← SE (small length scale)

Oh so... once we have the desired k, we can simply plug & chug to get the multivariate distribution for new points X_* (which can include X)

$$y_* \mid y, X, X_* \sim N(\bar{f}_*, \mathrm{cov}(f_*) + \sigma_n^2 I)$$

↑ also a parameter to estimate! → so include back in once found.

to get desired $k(x,x')$ Follow book example of RW:

$$k_1(x,x') = \theta_1^2 \exp\left(-\frac{(x-x')^2}{2\theta_2^2}\right) \rightarrow \text{long term smooth rising trend}$$

$$k_2(x,x') = \theta_3^2 \exp\left(-\frac{(x-x')^2}{2\theta_4^2} - \frac{2\sin^2(\pi(x-x'))}{\theta_5^2}\right) \rightarrow \text{seasonal variations}$$

θ_4: allow seasonal trend to decay away from periodicity

θ_5: smoothness of periodic component.

period = 1 (year). normally would be $\frac{2\sin^2\left(\frac{(x-x')}{2}\right)}{l^2}$

$\theta_8, \theta_7 > 0$

$$k_3(x,x') = \theta_6^2 \left(1 + \frac{(x-x')^2}{2\theta_8\theta_7^2}\right)^{-\theta_8}$$

→ (small) medium-term irregularities

amplitude

shape parameter
For diffuseness of length-scales

typical length-scale

→ rational quadratic term

(but radial quad. term better)

NOTE: could also use SE term for medium term irregularities

$$k_4(x,x') = \theta_9^2 \exp\left(-\frac{(x-x')^2}{2\theta_{10}^2}\right)$$

→ short-term ~~noise~~ effects

$$k_5(x,x') = \theta_{11}^2 I_n$$

→ independent noise of measurement...

n = # training data

SO Find best θ for our sample data (train)

$p(y|X,\theta)$

$\hookrightarrow y|X \sim N(y|0, C(\theta) + \theta_{11}^2 I_n)$, $C(\theta) = k_1(x,x') + k_2(x,x') + k_3(x,x') + k_4(x,x')$

$C(\theta) = k(x,x')$

0-mean prior } → means have to subtract mean(y) before do analysis (!)

$$\rightarrow \log(p(y|X,\theta)) = \log N(0, C(\theta) + \theta_{11}^2 I_n)$$

$$= \log\left[(2\pi)^{-n/2} |C(\theta) + \theta_{11}^2 I_n|^{-1/2} \exp\left(-\frac{1}{2}(C(\theta) + \theta_{11}^2 I_n)^{-1} y\right)\right]$$

$$= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|C(\theta) + \theta_{11}^2 I_n| - \frac{1}{2} y^T (C(\theta) + \theta_{11}^2 I_n)^{-1} y$$

gradient →

n = # of data points

now find ~~minima~~ maxima, using partial derivatives w.r.t. θ_j's

$$\frac{\partial}{\partial \theta_j} \log p(y|x,\theta) = \frac{1}{2} y^T K^{-1} \frac{\partial K}{\partial \theta_j} K^{-1} y - \frac{1}{2} \mathrm{tr}\left(K^{-1} \frac{\partial K}{\partial \theta_j}\right)$$

$$= \frac{1}{2} \mathrm{tr}\left(\left(\alpha\alpha^T - K^{-1}\right) \frac{\partial K}{\partial \theta_j}\right), \quad \alpha = K^{-1} y$$

$\alpha\alpha^T = K^{-1} y y^T K^{-1}$

(will simplify a lot...?)

$K = K(x, x')$

optimum? "conjugate gradient optimizer"...

→ Frechet differentiable. w/ ~~Boundary Cond. for~~ $\theta_7, \theta_8 > 0$

$\theta_{10}, \theta_5, \theta_4, \theta_2 \neq 0$

use vpa solve? w/ get where all gradients = 0... yeah, d guess so...

want $\frac{\partial}{\partial \theta_j} \log p(y|x,\theta) = 0$ @ (min/max) → iteratively solve d then compare values of multiple solutions → from different starting points (local maxima)

→ once we have this, plug back in to find $\bar{f}_*$, var(f_*) → draw from multivariate gaussian & plot!

Next step: 1) write up $K(x,x')$ in python & get expressions for ~~gradient~~

2) ~~attempt MATLAB solve for~~ $\nabla_\theta \log p(y|X,\theta) = 0$

3) ask prof (~~what minimize gradient??~~) 4) plug in θ solved & plot

$$\frac{\partial K}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}\begin{bmatrix} k_{11}(\theta) & k_{12}(\theta) & \cdots & k_{1n}(\theta) \\ k_{21}(\theta) & \ddots & & \vdots \\ \vdots & & \ddots & \\ k_{n1}(\theta) & \cdots & & k_{nn}(\theta) \end{bmatrix}$$

exp!

$$\text{if } K(x,x') = \sigma^2 \exp\left(-\frac{(x-x')^T(x-x')}{2l^2}\right)$$

$$\frac{\partial K}{\partial \sigma} = 2\sigma \exp(\ldots)$$

$$\frac{\partial K}{\partial l} = \sigma^2 \exp(\ldots)\ \frac{(x-x')^T(x-x')}{l^3}, \text{ yep}$$

$\frac{\partial}{\partial \theta_j}$ $\nabla_\theta \log p(y|X,\theta) = 0$ when each derivative $= 0$...
also at other locations... have to calculate
for values of θ though.

So... we could generate a bunch of samples & then
find local maxes & compare.
↳ need good starting points though

Gradient Descent for non-linear systems auto-choose α (or σ as it may be)
0.0001 → 13
logspace, 100 steps

steps:

~~1) post-together notebook structure~~

~~2) compare w/ nominal θ (& plot?)~~ → enforce rounding eig-values to $\frac{\delta}{\epsilon}$
(our eig-vectors & eig-vals round eig-vals to $\epsilon = 10$... should be good...

3) setup optimization for θ
↳ check params match

Gradient of kernel derivation

$$K_1(x,x') = \theta_1^2 \exp\left(\frac{-(x-x')^2}{2\theta_2^2}\right)$$

$$\frac{\partial K_1}{\partial \theta_1} = 2\theta_1 \exp(\dots)$$

$$\frac{\partial K_1}{\partial \theta_2} = \theta_1^2 \exp(\dots)(-2)\left(\frac{-(x-x')^2}{2\theta_2^3}\right)$$

$$= 2\theta_1^2 \exp\left(\frac{-(x-x')^2}{2\theta_2^2}\left(\frac{1}{2\theta_2^2} - \frac{1}{2\theta_2^3}\right)\right)$$

$$= 2\theta_1^2 \exp\left(\frac{-(x-x')^2}{2\theta_2^2}\right)\left(\frac{(x-x')^2}{2\theta_2^3}\right) \quad \checkmark$$

$$K_2(x,x') = \theta_3^2 \exp\left(\frac{-(x-x')^2}{2\theta_4^2} - \frac{2\sin^2(\pi(x-x'))}{\theta_5^2}\right)$$

$$\frac{\partial K_2}{\partial \theta_3} = 2\theta_3 \exp(\dots) \quad \checkmark$$

$$\frac{\partial K_2}{\partial \theta_4} = \theta_3^2 \exp(\dots)\left(\frac{(x-x')^2}{\theta_4^3}\right) \quad \checkmark$$

$$\frac{\partial K_2}{\partial \theta_5} = \theta_3^2 \exp(\dots)\left(\frac{2\sin^2(\pi(x-x'))}{\theta_5^3}\right)(2) \quad \checkmark$$

$$k_3(x,x') = \theta_6^2\left(1+\frac{(x-x')^2}{2\theta_7^2\theta_8}\right)^{-\theta_8}$$

$$\frac{\partial k_3}{\partial \theta_6} = 2\theta_6(\dots) \quad \checkmark$$

$$\frac{\partial k_3}{\partial \theta_7} = \theta_6^2\left(1+\frac{(x-x')^2}{2\theta_7^2\theta_8}\right)^{-\theta_8-1}\left(\frac{\theta_8(x-x')^2(-2)}{2\theta_7^3\theta_8}\right)$$

$$= \theta_6^2\left(1+\frac{(x-x')^2}{2\theta_7^2\theta_8}\right)^{-\theta_8-1}\left(\frac{(x-x')^2}{\theta_7^3}\right) \quad \checkmark$$

$$\frac{\partial k_3}{\partial \theta_8} = \theta_6^2\frac{\partial}{\partial\theta_8}\left(1+\frac{(x-x')^2}{2\theta_7^2\theta_8}\right)^{-\theta_8}$$

$$= \theta_6^2\frac{\partial}{\partial\theta_8}\exp\left(\log\left(1+\frac{(x-x')^2}{2\theta_7^2\theta_8}\right)^{-\theta_8}\right)$$

let $u = -\theta_8\log(\dots)$, $\frac{\partial}{\partial u}e^u = e^u\frac{\partial u}{\partial u}$, $\frac{\partial}{\partial\theta_8} = \frac{\partial}{\partial u}\frac{\partial u}{\partial\theta_8} \Rightarrow \frac{\partial e^u}{\partial\theta_8} = \frac{\partial e^u}{\partial u}\frac{\partial u}{\partial\theta_8}$

$$\Rightarrow \frac{\partial u}{\partial\theta_8} = -\log(\dots) + \left(-\theta_8\left(\frac{1}{1+\frac{(x-x')^2}{2\theta_7^2\theta_8}}\right)\left(\frac{(x-x')^2(-1)}{2\theta_7^2\theta_8^2}\right)\right)$$

$$\Rightarrow \frac{\partial k_3}{\partial\theta_8} = \theta_6^2\left[\exp\left(\log\left(1+\frac{(x-x')^2}{2\theta_7^2\theta_8}\right)^{-\theta_8}\right)\right]\left[-\log\left(1+\frac{(x-x')^2}{2\theta_7^2\theta_8}\right)+\left(\frac{(x-x')^2}{2\theta_7^2\theta_8+(x-x')^2}\right)\right]$$

$$= \theta_6^2\left[1+\frac{(x-x')^2}{2\theta_7^2\theta_8}\right]^{-\theta_8}\left[\frac{(x-x')^2}{2\theta_7^2\theta_8+(x-x')^2}-\log\left(1+\frac{(x-x')^2}{2\theta_7^2\theta_8}\right)\right] \quad \checkmark$$

11

$$k_4 = \theta_9^2 \exp\left(-\frac{(x-x')^2}{2\theta_{10}^2}\right)$$

$$\frac{\partial k_4}{\partial \theta_9} = 2\theta_9 \exp(\dots)$$

$$\frac{\partial k_4}{\partial \theta_{10}} = \theta_9^2 \exp(\dots)\left(\frac{(x-x')^2}{\theta_{10}^3}\right) \quad \checkmark$$

$$k_5 = \theta_{11}^2 I_n$$

$$\frac{\partial k_5}{\partial \theta_{11}} = 2\theta_{11}$$

$\log p(\theta)$. if $p(\theta)$ is a multivariate gaussian distribution, where each have their own mean & cov?

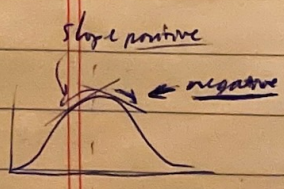
or is it like $\frac{p(\theta_1)+p(\theta_2)+p(\theta_3)+\dots+p(\theta_{11})}{11}$ ← (so that integrates to 1)

if Gaussian, then $p(\theta_1) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2}\frac{(\theta_1-\mu_1)^2}{\sigma_1^2}\right)$

$$\rightarrow \log p(\theta_1) = -\frac{1}{2}\log(2\pi) + 2\log\sigma_1 + \left(-\frac{1}{2}\frac{(\theta_1-\mu_1)^2}{\sigma_1^2}\right)$$

$$\frac{\partial}{\partial\theta_1}\log p(\theta_1) = 0 + 0 - \frac{1}{2}\frac{(\theta_1-\mu_1)}{\sigma_1^2}(2)$$

$$= \frac{\mu_1 - \theta_1}{\sigma_1^2}$$



if $b_1 \sim (0,1)$, $\mathbb{E}[b_1 b_1^T] = 1$

$b_2 \sim (0,2)$, $\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \sim ?$

$$\mathbb{E}\left[\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}\right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\mathbb{E}\left[\begin{pmatrix} b_1 \\ b_2 \end{pmatrix}\begin{pmatrix} b_1^T & b_2^T \end{pmatrix}\right] = \mathbb{E}\begin{bmatrix} b_1 b_1^T & b_1 b_2^T \\ b_2 b_1^T & b_2 b_2^T \end{bmatrix} = \begin{bmatrix} \sigma_1 & \mathbb{E}[b_1 b_2^T] \\ & \sigma_2 \end{bmatrix}$$

$$\mathbb{E}[b_1 b_2^T] = \int b_1 b_2^T f_{b_1 b_2}(b_1, b_2)\, db_1\, db_2$$

Yeah... hard to tell/define w/ our knowing distribution joint.

→ if assume independent, then $=0$

→ for explicit basis function... $[1 \quad x \quad x^2 \quad x^3 \quad x^4]$

but then be independent w/ defined means + hope get us in ball park...

↳ normalize this data set?? to make it easier for computing weights?