Recap

- $\underset{S}{\text{Rad}}(\mathcal{H}) = \underset{\sigma}{\mathbb{E}} \sup_{h \in \mathcal{H}} \sigma \cdot \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_m) \end{bmatrix}$

Rademacher complexity

$\sigma = [\sigma_1, \ldots, \sigma_m]^T \in \{-1, 1\}^m$

with Pr $1/2$, $\sigma_i = 1$ iid

$S = \{(x_i, y_i)\}_{i \in [m]}$

- Several notions of function complexity $\nearrow R(h)$

$$\text{generalization gap} = \overbrace{\underset{(x,y)\sim\mathcal{D}}{\mathbb{E}} \ell(x, y, h(x))} - \underbrace{\frac{1}{m}\sum_{i=1}^{m} \ell(x_i, y_i, h(x_i))}_{\hat{R}_S(h)}$$

$$\leq \underset{S}{\text{Rad}}(\mathcal{H}) + O\left(\sqrt{\frac{\log 1/\delta}{2m}}\right)$$

with Pr $\geq 1-\delta$ (over $S \sim \mathcal{D}^m$)

- $\underset{S}{\text{Rad}}(\mathcal{H}) \leq \frac{\Lambda \|G\|_F}{m}$

RKHS

(Proof : last time)

$\Lambda = \sup_{\alpha} \|\alpha\|$

m : # samples

G : Gram matrix evaluated on S.

- kernel regression : why is it effective?
- Regularization term : $Y^{*}Y$ and its effect on Fourier coefficients of $f \in \mathcal{H}$.

- Complexity of $\mathcal{H}$ $\quad$ ($Rad_S(\mathcal{H})$)

- $x \to \mathcal{K}(x, \cdot)$ $\quad$ (RKHS map)

  $x \to \{\sqrt{\lambda_j}\ \psi_j(x)\}_j$ $\quad$ (Mercer map)

  dimension of feature ~~space~~

  $\iff$ dim of kernel regression

  small when $\lambda_j \to 0$

Bayesian methods

Gaussian process regression

Want to learn h s.t. $h(x_i) \approx y_i$
$i \in [m]$

$P(h \mid X, y)$

, what is the probab. of finding h given $S = \{(x_i, y_i)\}_{i \in [m]}$

- Beyond pointwise prediction, want uncertainties
- logistic regression

Bayes' rule

$$P(h \mid X, y) = \frac{\text{Prior}(h) \times P(X, y \mid h)}{P(X, y)}$$

$P(X, y \mid h)$: likelihood
model evidence

$$P(X, y) = \int P(X, y \mid h)\, dP\text{rior}(h)$$

- $X, y = \{(x_i, y_i)\}_{i=1}^m$ iid

$$P(X, y \mid h) = \prod_{i=1}^m P(y_i - h(x_i))$$

- how to choose a prior on h?

Gaussian process: is a function $x \to h(x)$ s.t. for every m, $\{h(x_1), h(x_2), .., h(x_m)\}$ is multivariate normally distributed.

$$\begin{bmatrix} h(x_1) \\ h(x_2) \end{bmatrix} \sim \frac{1}{2\pi \sqrt{\det C}} e^{-[a\ b]^T C^{-1} [a, b]}$$

$a = h(x_1) - \mu_1$
$b = h(x_2) - \mu_2$

$$\begin{bmatrix} h(x_1) \\ \vdots \\ h(x_m) \end{bmatrix} \sim \frac{1}{(2\pi)^{m/2} \sqrt{\det C_m}} e^{-[a_1 \dots a_m]^T C_m^{-1} [a_1, \dots, a_m]}$$

Ex. C_m is PD.

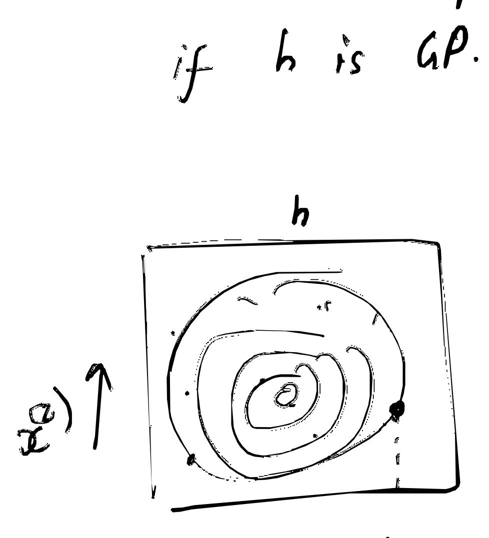
$$C_{ij} = \mathbb{E}(h(x_1) h(x_2)) - \mathbb{E} h(x_1) \mathbb{E} h(x_2)$$

Want to sample $S = \{(x_i, y_i)\}_{i=1}^m$

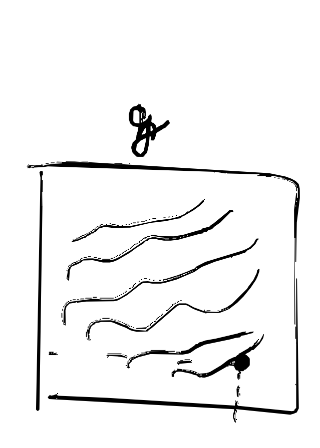
Bayes' rule

$$P([h(x_1'), h(x_2') \dots, h(x_m')] \mid S) = \frac{1}{P(S)} P(S \mid h(x_1') \dots, h(x_m')) \times P(h(x_1'), \dots h(x_m'))$$

Under Gaussian process assumption
if h is GP.



$x_i = [x^{(1)}(x_i)\ \ x^{(2)}(x_i)]$



$(x, y) \sim \mathcal{D}$

Posterior

$$P([h(x_1) \dots h(x_m)] \mid S) = \frac{1}{P(S)} P(S \mid h(x_1) \dots h(x_m)) \times P(h(x_1) \dots h(x_m))$$

$\mathcal{D}^m$

Under GP assumption

$$P(h(x_1), \dots h(x_m)) = \frac{e^{-(h(x_1) \dots h(x_m))^T C_m^{-1} \begin{pmatrix} h(x_1) \\ \vdots \\ h(x_m) \end{pmatrix}}}{(2\pi)^{m/2} \sqrt{\det C_m}}$$

e.g. $C(h(x_1), h(x_2)) = \mathcal{K}(x_1, x_2)$

Where is the posterior maximized?
: Bayesian view of regression

what is the value of $h(x)$ at a new x?
(pointwise regression)

what is $\max_h P(h(x) \mid S)$?

$$P(h(x) \mid S) = \frac{P(S \mid h)\, P(h)}{P(S)}$$

$$P\left(\begin{pmatrix} h(x_1) \\ \vdots \\ h(x_m) \end{pmatrix}\right) = \text{M.V.G.}$$

$$P(S \mid h) = \prod_{i=1}^m P(y_i - h(x_i))$$

e.g. Gaussian likelihood

$y_i = h(x_i) + \varepsilon_i$
(labels / observations)
$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

$$\max_h P(h \mid S) = \max_h P(S \mid h)\, P(h)$$

$\operatorname{argmax}_h P(h(x) \mid S)$: Maximum a posteriori estimate
MAP

- log ↑.

$$\Rightarrow \operatorname{arg\,max}_h P(h \mid S) = \operatorname{argmax}_h \log P(h \mid S)$$

$$= \operatorname{argmax}_h \log P(S \mid h) + \log P(h)$$

Under GP assumption for prior and Gaussian assumption on likelihood,

$$= \operatorname{argmax}_h - \sum_{i=1}^m \frac{(y_i - h(x_i))^2}{2\sigma^2} - (h(x_1) \dots h(x_m))^T G^{-1} \begin{pmatrix} h(x_1) \\ \vdots \\ h(x_m) \end{pmatrix}$$

$$= \operatorname{argmin}_{h \in \mathcal{H}} \sum_{i=1}^m \frac{(y_i - h(x_i))^2}{2\sigma^2} + (h(x_1) \dots h(x_m))^T G^{-1} \begin{pmatrix} h(x_1) \\ \vdots \\ h(x_m) \end{pmatrix}$$

Use representer theorem:

$$h(x) = \sum_{i=1}^m \alpha_i \mathcal{K}(x_i, x)$$

$$\Rightarrow \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_m) \end{bmatrix} = G\alpha \qquad \alpha \in \mathbb{R}^m$$

Applying same old kernel trick

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2\sigma^2} \| Y - G\alpha \|^2 + \alpha^T G \alpha$$

$Y = [y_1, \dots, y_m]^T$

$$\alpha = (G + \sigma^2 I)^{-1} Y$$

Bayesian view → automatic regularization

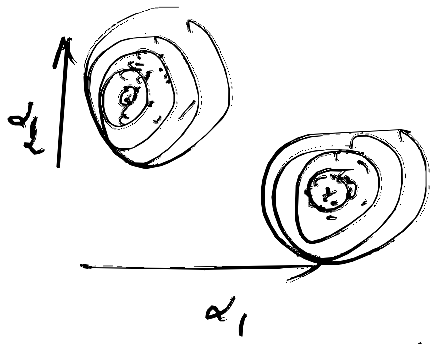
Solving for MAP estimate.

Now want: samples from Bayesian posterior

- Bayesian inference:

$$\text{Posterior}(\alpha \mid S) = \frac{\text{Prior}(\alpha) \times P(S \mid \alpha)}{P(S)}$$

$$\nabla \log \text{Posterior}(\alpha \mid S) = \nabla \log \text{prior}(\alpha) + \nabla \log P(S \mid \alpha)$$



Score function:
$\nabla \log$ density $: \mathcal{X} \to \mathbb{R}^d$

Setting of B.I: (typically)
Score function of target: known

- Settings for sampling from target distributions

Generative modeling:
Want samples from prob. dist. π.

Given: $\alpha_1, \ldots, \alpha_m \sim \pi$.
m samples

Algorithms for sampling and generative modeling that use kernels.

Generative modeling: GAN
MMD GAN

Sampling: Stein Variational gradient descent

Integral probability metrics

P, Q are two probabi. distributions

$Q, P: \mathcal{X} \to \mathbb{R}^{+}$

$$d_{\mathcal{F}}(P,Q) = \sup_{f \in \mathcal{F}} \left| \underset{X \sim P}{\mathbb{E}} f(X) - \underset{X \sim Q}{\mathbb{E}} f(X) \right|$$

- $\mathcal{F}$: Lipschitz functions with Lipschitz constant 1.

i.e.

$$\| f(x) - f(y) \| \leq \| x - y \| \quad \forall f \in \mathcal{F}.$$

$d_{\mathcal{F}}(P,Q)$: Wasserstein-1 distance

- $\mathcal{F}$: RKHS and $\|f\|_{\mathcal{H}} \leq 1$

$d_{\mathcal{F}}(P,Q) :=$ Maximum mean discrepancy