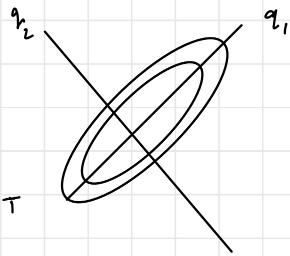
Lecture 12

- Gaussian Distribution:

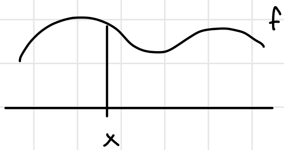


$$\begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \end{pmatrix}, \begin{array}{cc} & j \\ i & \boxed{\Sigma_{ij}} \end{array} \right), \quad \Sigma = Q \Lambda Q^T$$

$$\Sigma \geq 0$$

finite index

- Gaussian Process:



noiseless $y = f(x)$

$$y = \cdot x \begin{pmatrix} y_x \end{pmatrix} = \begin{pmatrix} f(x) \end{pmatrix} = f \sim GP\left(\boxed{0}, \begin{array}{cc} & x' \\ x & \boxed{\sigma^2 K(x,x')} \end{array} \right)$$

continuous index

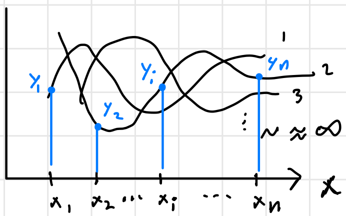
$$K \geq 0$$

$K = Q \Lambda Q^T$ Mercer Thm.

$$\begin{array}{cc} & x' \\ x & \boxed{K(x,x')} \end{array} = \begin{array}{cc} & 1 \; 2 \cdots k \cdots \\ x & \boxed{} \end{array} \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_k \end{pmatrix} \begin{array}{c} x' \\ \boxed{} \end{array}$$

- Prior : $f \sim GP(0, \sigma^2 K)$

- Population :

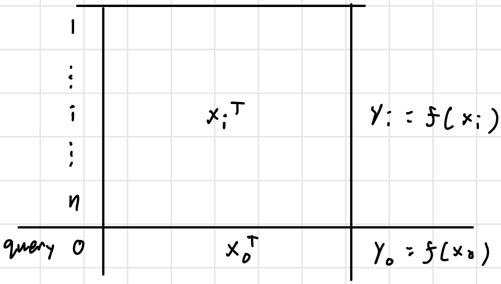


$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i = f(x_i) \\ \vdots \\ y_n \end{pmatrix} \sim \mathcal{N}\left(\begin{array}{c} 1 \\ 2 \\ \vdots \\ i \\ \vdots \\ n \end{array} \begin{bmatrix} \\ \\ \\ 0 \\ \\ \\ \end{bmatrix}, \begin{array}{cc} & \begin{array}{ccccc} 1 & 2 \cdots & j & \cdots & n \end{array} \\ \begin{array}{c} 1 \\ 2 \\ \vdots \\ i \\ \vdots \\ n \end{array} & \begin{bmatrix} & & \\ & & \\ & & \\ & \sigma^2 K_{ij} & \\ & & \\ & & \end{bmatrix} \end{array} \right)$$

$$K_{ij} = K(x_i, x_j)$$

- Gaussian Process for Learning:

Training Data



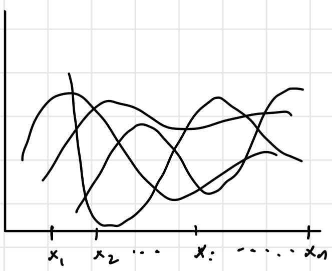
- Bayesian Inference:

$$[f \mid f(x_i) = y_i, \ i = 1, \dots, n]$$

Among $N \approx \infty$ curves

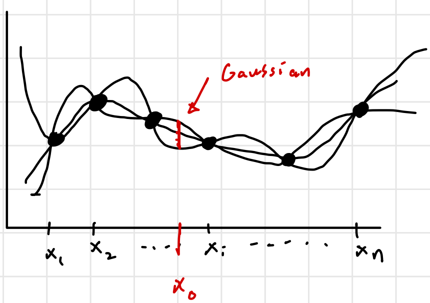
$M \ll N$ curves with $f(x_i) = y_i, \quad i = 1, \dots, n$

Prior



$\mapsto$

Posterior



(e.g. temperatures @ different locations)
spatial statistics

"Learning is memory + interpolation"

$$
Y \left[\begin{array}{c} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \\ \hline y_0 \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} 0 \\ \\ \\ \hline \\ \end{array}\right], \left[\begin{array}{c|c} \tau^2 K_{n \times n} & \tau^2 k_0^T \\ \hline \tau^2 k_0 & k_{00} \end{array}\right] \right)
$$

$$k_{0j} = K(x_0, x_j)$$

So

$$[y_0 \mid y_1, \dots, y_n] \sim \mathcal{N}\big(\tau^2 k_0, \ (\tau^2 K \underset{+\sigma^2 I_n}{\uparrow})^{-1} Y, \ \text{variance}\big)$$

$+\sigma^2 I_n$

if noise $\quad y_i = f(x_i) + \varepsilon_i$

- $$\mathbb{E}(y_0 \mid Y) = \tau^2 k_0 (\tau^2 K + \sigma^2 I_n)^{-1} Y$$

$$= k_0 \Big(K + \underbrace{\boxed{\tfrac{\sigma^2}{\tau^2}}}_{\lambda} I_n\Big)^{-1} Y$$

$$= k_0 \underbrace{(K + \lambda I_n)^{-1}}_{C}$$

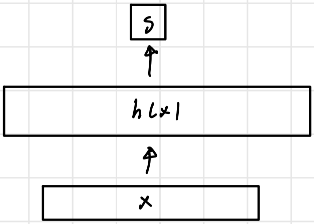
$$= \left[\begin{array}{c} K(x_i, x_0) \end{array}\right] \left[\begin{array}{c} c_i \end{array}\right]$$

$$\hat{f}(x_0) = \sum_{i=1}^{n} c_i K(x_i, x_0)$$

- Another Interpretation :

$$\hat{f}(x_0) = \sum_{i=1}^{n} \underbrace{c_i}_{\text{value}} \overbrace{K(\underbrace{x_i}_{\text{key}}, \underbrace{x_0}_{\text{query}})}^{\text{affinity}}$$

- Supervised Learning :



$s = f(x) = h(x)^T \beta$

Kernel regression :

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - h(x_i)^T \beta)^2 + \lambda |\beta|^2$$

Gaussian Process :

$$[\beta \mid Y, X] \propto p(\beta) \prod p(Y_i \mid x_i, \beta)$$

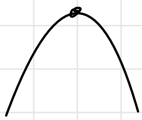
$$\propto \exp\left(-\frac{1}{2\tau^2}|\beta|^2 + \frac{1}{2\sigma^2}(Y_i - h(x_i)^T\beta)^2\right)$$

$$-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(Y_i - h(x_i)^T\beta)^2 + \underbrace{\frac{\sigma^2}{\tau^2}}_{\lambda}|\beta|^2\right)$$

Gives a posterior mode $\hat{\beta}$

$\longrightarrow f(x_0) = h(x_0)^T \hat{\beta}$

$P(\beta \mid X, Y)$



But for the Gaussian Distribution, the posterior expectation is the posterior mode. $E[\beta \mid X, Y] = \hat{\beta}$

$$E(f(x_0) = h(x)^T\beta \mid X, Y) = \hat{f}(x_0) = h(x_0)^T \hat{\beta}$$

- Advantage of GP:

(1) Variance V

$$[y_0 \mid Y] \sim N(\quad, V)$$

uncertainty quantification

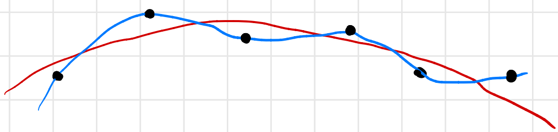
(2) hyper-parameter tuning

$$K(x, x') = \exp(-\gamma |x - x'|^2), \quad \tau^2, \sigma^2$$

$$Y \sim N(0, \overbrace{\tau^2 K_\gamma + \sigma^2 I}^{\Sigma})$$

$$\text{Likelihood}(\tau^2, \gamma, \sigma^2) = p(Y)$$

$$= \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2} Y^T \Sigma^{-1} Y\right)$$



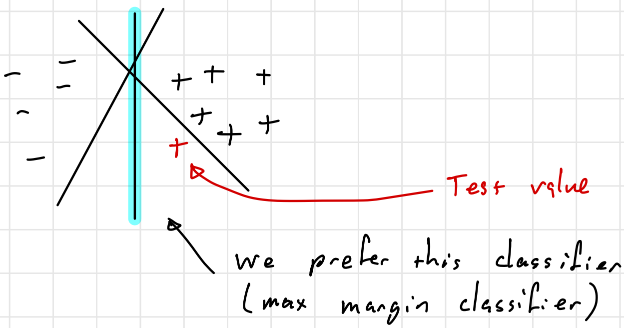
- Support Vector Machine (SVM)

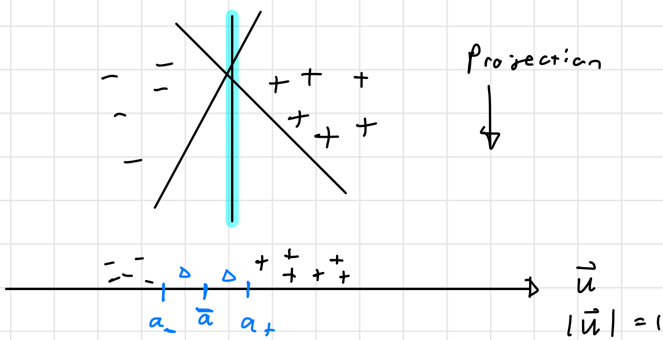
  - binary classification

$$s_i = \langle x_i, w \rangle + b$$

perceptron: $\hat{y}_i = \text{sign}(s_i) = \begin{cases} + & \text{if } s_i \geq 0 \\ - & \text{if } s_i < 0 \end{cases}$

- Geometry (seperable):





$\langle x_i, \vec{u} \rangle$ projection on $\vec{u}$

Let $a_+ = \min\limits_{i: y_i = +} \langle x_i, \vec{u} \rangle$, $a_- = \max\limits_{i: y_i = -} \langle x_i, \vec{u} \rangle$

$$\bar{a} = \frac{a_+ - a_-}{2} \quad \& \quad \Delta = a_+ - \bar{a} = \bar{a} - a_-$$

- We need to find $\vec{u}$ that maximizes Δ

$$\langle x_i, \vec{u} \rangle \geq a_+ \quad \text{for} \quad y_i = +1$$

$$\langle x_i, \vec{u} \rangle \leq a_- \quad \text{for} \quad y_i = -1$$

" = " attainable.

- translate the optimization problem.

$$\langle x_i, \vec{u} \rangle \geq a_+ \quad \text{for} \quad y_i = +1 \Rightarrow \langle x_i, \vec{u} \rangle - \bar{a} \geq \Delta$$

$$\langle x_i, \vec{u} \rangle \leq a_- \quad \text{for} \quad y_i = -1 \Rightarrow \langle x_i, \vec{u} \rangle - \bar{a} \leq -\Delta$$

So,

$$\langle x_i, \overset{w}{\boxed{\frac{\vec{u}}{\Delta}}} \rangle \overset{b}{\boxed{- \frac{\bar{a}}{\Delta}}} \geq 1 \quad \text{for} \quad y_i = +1$$

$$\langle x_i, \frac{\vec{u}}{\Delta} \rangle - \frac{\bar{a}}{\Delta} \leq -1 \quad \text{for} \quad y_i = -1$$

$$\left.\begin{aligned} \langle x_i, w \rangle + b &\geq 1 \quad \text{for} \quad y_i = +1 \\ \langle x_i, w \rangle + b &\leq 1 \quad \text{for} \quad y_i = -1 \end{aligned}\right\} \rightarrow y_i s_i = y_i (\langle x_i, w \rangle + b) \geq 1$$

Note: $|w| = \left|\frac{\vec{u}}{\Delta}\right| = \frac{1}{\Delta}$

- So $\Delta = \frac{1}{|w|}$

- To maximize Δ we want:

$$\boxed{\begin{aligned} &\min \ \tfrac{1}{2}|w|^2 \\ &\text{s.t.} \quad y_i(\langle x_i, w\rangle + b) \geq 1 \\ &\qquad\qquad i = 1, \dots, n \end{aligned}}$$

Optimization Problem

- Lagrangian:

$$L(\underbrace{(w, b)}_{\text{primal}}, \underbrace{\alpha}_{\text{dual}}) = \frac{1}{2}|w|^2 + \sum_{i=1}^{n} \alpha_i (1 - y_i(\langle x_i, w\rangle + b))$$

Primal Problem

$$\alpha = \begin{pmatrix} \alpha_i \geq 0 \end{pmatrix}_1^n$$

$$\min_{(w,b)} \max_{\alpha > 0} L((w,b), \alpha)$$

new constraint, much simpler

Claim: all old constraints $y_i(\langle x_i, w\rangle + b) \geq 1$ $i = 1 \cdots n$ will be automatically satisfied

Proof by contradiction:

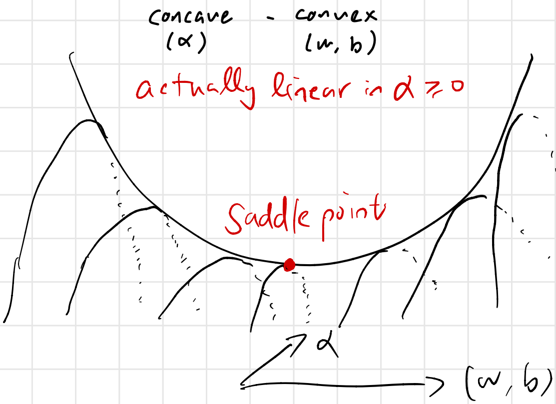
Otherwise $\exists i$ s.t. $y_i(\langle x_i, w\rangle + b) < 1$

$$\max_{\alpha_i > 0} \underbrace{\alpha_i}_{>0} \left(\underbrace{1 - y_i(\langle x_i, w\rangle + b)}_{>0}\right) \longrightarrow \infty$$

$\downarrow$

∞

can not be $\min_{(w,b)}$!



Saddle point

$$\min_{(w,b)} \max_{\alpha \geq 0} = \max_{\alpha \geq 0} \min_{(w,b)}$$

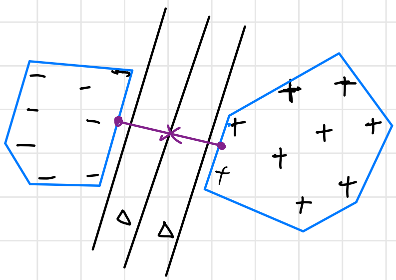
Von Neumann Zero sum game

$$\max_{\alpha \geq 0} \boxed{\min_{(w,b)} \mathcal{L}((w,b),\alpha)}$$

↓ closed form, representer

$$\max_{\alpha \geq 0} Q(\alpha)$$ Dual Problem

- Geometry :



convex hulls

primal : max margin

dual : min distance

- non - seperable : slackness

relax to $\sum_{i=1}^{n} \xi_i$

$$\min \tfrac{1}{2} |w|^2 + C \, \# \text{ of } \xi_i > 0$$ → hinge loss

$$\text{s.t. } \langle x_i, w \rangle + b \geq 1 - \xi_i, \quad i = 1, \dots, n$$