ECE 544 NA

10/14/2013

EXAM 1: 11/12, 11:00 - 12:20, 1068 ENGINEERING HALL

TODAY: RBF

RECALL: GENERALIZED LINEAR CLASSIFIER

$$y_k(x^n) = \sum_{j=0}^{M} w_{kj}\,\phi_j(x^n) \qquad \phi_0(x) \equiv 1$$

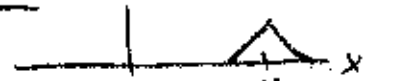
$\phi_j(x^n)$ = "BASIS FUNCTION"

RADIAL BASIS FUNCTION $\equiv$ A BASIS FUNCTION S.T.

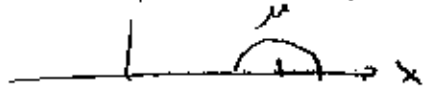
$$\phi_j(x^n) = \phi\left(\|x^n - \mu_j\|\right) \quad \text{FOR SOME } \mu_j$$

① ENGINEER CHOOSES ϕ BASED ON INTUITION

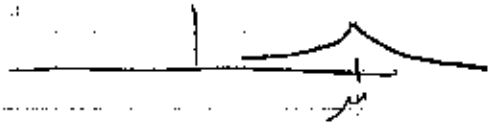
EXAMPLES $\phi(x) = \max(0,\ 1 - \lambda\|x-\mu\|)$



$\phi(x) = \max\left(0,\ 1 - \frac{\|x-\mu\|^2}{\sigma^2}\right)$



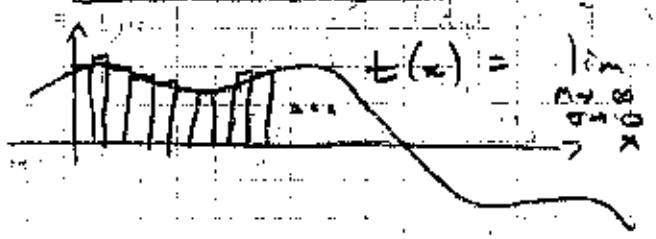
$\phi(x) = e^{-\lambda\|x-\mu\|}$



$\phi(x) = e^{-\frac{1}{2}\frac{\|x-\mu\|^2}{\sigma^2}}$ ← MOST COMMON

DIGRESSION: ANY TARGET FUNCTION $t(x)$ CAN BE WRITTEN $t(x) = \lim_{\substack{M\to\infty \\ \sigma\to 0}} \sum_{j=0}^{M} w_{kj}\,\phi_j(x)$

PROOF: NEWTON'S INTEGRAL THEOREM



$$t(x) = \lim_{\substack{M\to\infty \\ \sigma\to 0}} \sum_{j=1}^{M} t(\mu_j)\,u\left(\sigma - \|\mu_j - x\|\right)$$

(2) GIVEN Φ, COMPUTER FIND $\vec{\mu}_j$, w_{kj}

E.G.

$$\vec{\mu}_j, w_{kj} = \operatorname{argmin} \sum_{n=1}^{N} \sum_{k=1}^{C} \left(t_k^n - y_k^n \right)^2$$

(a) FIND w_{kj} GIVEN FIXED μ_j; ANALYTIC SOLUTION EXISTS (PSEUDO-INVERSE)

$$w_{kj} = \operatorname{argmin} \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{C} \left(t_k^n - \sum_{j=0}^{M} w_{kj} \Phi_j(x^n) \right)^2$$

T $[N \times C]$ $\quad T(n,k) = t_k^n = [\vec{t}_1, \ldots, \vec{t}_C]$

Φ $[N \times (M+1)]$ $\quad \Phi(n, j+1) = \Phi_j(x^n)$

W $[C \times (M+1)]$ $\quad W(k, j+1) = w_{kj} = \begin{bmatrix} \vec{w}_1^T \\ \vdots \\ \vec{w}_C^T \end{bmatrix}$

$$E = \frac{1}{2} \sum_{k=1}^{C} \| \vec{t}_k - \Phi \vec{w}_k \|^2$$

$$\nabla_{\vec{w}_k} E = \Phi^T (\Phi \vec{w}_k^T - \vec{t}_k) = 0$$

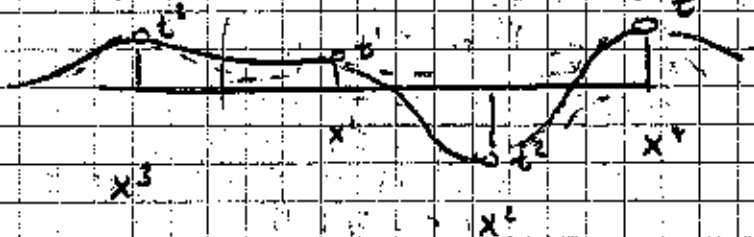
AT $\boxed{\vec{w}_k^T = (\Phi^T \Phi)^{-1} \Phi^T \vec{t}_k}$

$$\boxed{W^T = (\Phi^T \Phi)^{-1} \Phi^T T}$$

(b) FIND μ_j : THREE METHODS COMMON

(i) PARZEN WINDOWS / KERNEL ESTIMATOR

SET $\boxed{M = N, \ \vec{\mu}_j = \vec{x}^j}$; $y(x) = \sum_{j=1}^{N} \Phi(\|\vec{x} - \vec{x}^j\|)$



(ii) GRADIENT OPTIMIZATION

$$\mathcal{E} = \frac{1}{4}\sum_{n=1}^{N}\sum_{k=1}^{C}\left(t_k^n - \sum_{j=0}^{M} w_{kj}\,\phi(\|x^n - \mu_j\|^2)\right)^2$$

$$\nabla_{\mu_j}\mathcal{E} = \sum_{n=1}^{N}\left[\sum_{k=1}^{C}\left(\sum_{i=0}^{M}\left(w_{ki}\,\phi(\|x^n-\mu_i\|^2)\right) - t_k^n\right)\right]\phi'(\|x^n-\mu_j\|^2)\,(\mu_j - x^n)$$

CALL THIS $\gamma_j(x^n)$

⇒ EM-LIKE GRADIENT DESCENT ALGORITHM

① $$\mu_j = \frac{\sum_{n=1}^{N}\gamma_j(x^n)\,x^n}{\sum_{n=1}^{N}\gamma_j(x^n)}$$

② RECOMPUTE w_{kj}

③ RECOMPUTE $\gamma_j(x^n)$

④ GO TO ① UNTIL ALL CONVERGENCE

(iii) K-MEANS

$$\mathcal{E} = \frac{1}{2}\sum_{n=1}^{N}\min_j \|x^n - \mu_j\|^2$$

$$\nabla_{\mu_j}\mathcal{E} = \sum_{j=1}^{M}\sum_{n}\left[\left[j = j^*_n\right]\right](\mu_j - x_n)$$

① COMPUTE $j^*_n = \operatorname{argmin}_j \|x^n - \mu_j\|^2$

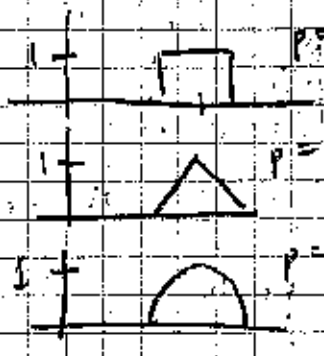
② FIND $\mu_j = \dfrac{\sum_n [j = j^*_n]\,x_n}{\sum_n [j = j^*_n]}$

③ REPEAT UNTIL CONVERGENCE

(3) How can we set ϕ_j, σ?

A: You can't! (not based on training data alone)

Proof: Let $\phi_j(\vec{x}^n - \vec{\mu}_j) = \dfrac{\sigma - \|\vec{x}^n - \vec{\mu}_j\|_p^p}{\sigma}$



etc.

Let $M = N$, $\vec{\mu}_j = \vec{x}^j$, $w_{kj} = t_k^j$

Then, for any small enough σ_j, for any p,

$$E = \sum_{n=1}^{N} \sum_{k=1}^{c} \left(t_k^n - \sum_{j=1}^{M} [j=n]\, t_k^j \cdot 1 \right)^2 = 0 !$$

Knowledge beyond the dataset: Regularization

$$E = \frac{1}{2} \sum_{n=1}^{N} \left(t^n - y(x^n) \right)^2 + \frac{\nu}{2} \int |P y(x)|^2 \, dx$$

- ν: penalty for non-smoothness
- $|P y(x)|^2$: estimate of non-smoothness of $y(x)$

$$\frac{\partial E}{\partial y(x)} = \begin{cases} \nu \hat{P} P y(x) & x \notin x^n \\ \nu \hat{P} P y(x) + \left(y(x) - t^n \right) \delta(x - x^n) & x \in x^n \end{cases}$$

$\hat{P} P y(x)$: How much does changing $y(x)$ at this particular x affect the average non-smoothness

SOLUTION: SETTING $\frac{\partial \mathcal{E}}{\partial y(x)} = 0$ AT EACH x

GIVES $$y(x) = \sum_{n=1}^{N} w_n G(x - x^n)$$

FOR $G(x)$ S.T. $\hat{P}\hat{P} G(x - x^n) = \delta(x - x^n)$

EXAMPLE

$$\mathcal{E} = \sum_{n=1}^{N} (y(x^n) - t^n)^2 + \underbrace{\sum_{l=0}^{\infty} \frac{\sigma^{2l}}{l! 2^l}}_{\text{NORMALIZE LIKE THE TAYLOR SERIES FOR } e^x} \int \underbrace{|\nabla^2 y(x)|^l}_{l^{\text{TH}} \text{ POWER OF CURVATURE}} dx$$

$$\frac{\partial \mathcal{E}}{\partial y(x)} = 0 \quad \text{AT EACH } x$$

$$\Rightarrow \quad y(x) = \sum_{n=1}^{N} w_n \, e^{-\frac{1}{2} \frac{\|x - x^n\|^2}{\sigma^2}}$$

ONE MORE PERSPECTIVE: REPRESENTER THEOREM

MERCER'S THEOREM

SUPPOSE $\Phi(x - x^n)$ IS A POSITIVE DEFINITE KERNEL, MEANING THAT FOR ANY $f(x)$,

$$\iint \Phi(x - x^n) f(x)^* f(x^n) \, dx \, dx^n > 0$$

THEN

$$\Phi(x - x^n) = \sum_{j=0}^{\infty} \lambda_j \Psi_j(x) \Psi_j(x^n)$$

$[\lambda_j, \Psi_j]$ = EIGENVALUES, EIGENFUNCTIONS OF Φ

AND THEREFORE L_2 NORMS EXIST:

$$\|\vec{\Psi}(x^n) - \vec{\Psi}(x^m)\|^2 = \|\vec{\Psi}(x^n)\|^2 + \|\vec{\Psi}(x^m)\|^2 - 2\vec{\Psi}(x^n)^T \vec{\Psi}(x^m)$$

$$= \sum_{j=0}^{\infty} \lambda_j \left(\Psi_j^2(x^n) + \Psi_j^2(x^m) - 2\Psi_j(x^n) \Psi_j(x^m) \right)$$

EXAMPLE

$$e^{-\|x-x^n\|^2} = 1 + (x-x^n)^T(x-x^n) + \frac{1}{2}\left[(x-x^n)^T(x-x^n)\right]^2 + \cdots$$

$$= \sum_{j=0}^{\infty} \psi_j(x)\,\psi_j(x^n)\,\lambda_j$$

$$\vec{\psi}(x^n) = \begin{bmatrix} 1 \\ x_1^n \\ \vdots \\ x_D^n \\ x_1^n x_1^n \\ x_1^n x_2^n \\ \vdots \\ x_D^n x_D^n \\ \vdots \end{bmatrix} \begin{matrix} \leftarrow 0^{TH}\ \text{ORDER} \\ \left.\begin{matrix} \\ \\ \end{matrix}\right\} \leftarrow 1^{ST}\ \text{ORDER} \\ \left.\begin{matrix} \\ \\ \\ \\ \end{matrix}\right\} 2^{ND}\ \text{ORDER} \\ \left.\begin{matrix} \\ \end{matrix}\right\} \text{ETC} \end{matrix} \qquad x^n = \begin{bmatrix} x_1^n \\ \vdots \\ x_D^n \end{bmatrix}$$

REPRESENTER THEOREM LET $y(x) = \sum_{j=0}^{\infty} w_j \psi_j(x) = \vec{w}^T \vec{\psi}(x)$

SUPPOSE $\vec{w}^* = \operatorname{argmin} \sum_{n=1}^{N} \left(t_n - \vec{w}^T \vec{\psi}(x^n)\right)^2 + \frac{\nu}{2}\|\vec{w}\|^2$

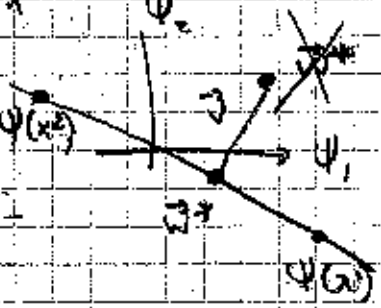
BUT IT'S ALWAYS TRUE THAT

$$\vec{w}^* = \underbrace{\sum_{n=1}^{N} \alpha_n \psi(x^n)}_{\text{IN THE SPAN OF } x^n} + \underbrace{\vec{v}}_{\text{ORTHOGONAL TO THE SPAN OF } x^n} \quad \text{FOR}$$

$$\|\vec{w}^*\|^2 = \left\|\sum_{n=1}^{N} \alpha_n \psi(x^n)\right\|^2 + \|\vec{v}\|^2$$

$$\sum_{n=1}^{N} \left(t^n - \vec{w}^T\vec{v} - \sum_{m=1}^{N} \alpha_m \vec{w}^T \psi(x^n)\right)^2 \leftarrow N \text{ UNKNOWNS}$$

CHOOSE TO MAKE THIS ZERO!



THUS $\vec{v}^{opt} = 0$, $\quad \vec{w}^* = \sum_{n=1}^{N} \alpha_n \psi(x^n)$

THEN $y(x) = \vec{w}^T\psi(x) = \sum_{n=1}^{N} \alpha_n \psi(x)^T \psi(x^n)$

$y(x) = \sum_{n=1}^{N} \alpha_n \Phi(x, x^n)$ MINIMIZES $\mathcal{E}$