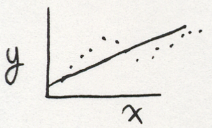
Feb 3, 2004

I <u>Regression</u>: predicting y from x $\quad p(y|x)$



- For simplicity, first assume a linear predictor

$$y = a^T x + b + \underbrace{\varepsilon}_{\text{gaussian noise } N(0,\Sigma)}$$

- To be fancy, let y be a linear sum of basis fns

$$y = \sum_i a_i \phi_i(x) + \varepsilon$$

- We find a_i by least squares. Let $a = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$

$$\hat{a} = \arg\min_a \sum_i \frac{\left[y_i - \sum_j a_j \phi_j(x_i)\right]^2}{2\sigma^2}$$

} we find set of coefficients that minimizes squared error between actual $\{y_i\}$'s and <u>predicted</u> $\{\tilde{y}_i\}$'s

<u>Aside</u>: A common problem is that the input ~~dimensions~~ ~~basis dimensions that are~~ x_i's are correlated ~~in correlated~~

This is often the case when n is large ("~~range~~ curse of dimensionality")

ie. Consider the area of the surface of a cube in n dimensions as the difference of the volumes of 2 cubes, one inside the other:

$$\text{volume} = \underbrace{1^n}_{\text{outer cube}} - \underbrace{(1-\varepsilon)^n}_{\text{inner cube}} \longrightarrow 1 \text{ as } n \to \infty.$$

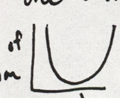
$\therefore$ all volume is located on skin of cube

$\therefore$ points randomly distributed in cube will tend to lie on surface & have correlations

~~problem: if the components of x are correlated~~
~~then $x_1, \ldots, x_K$ and~~

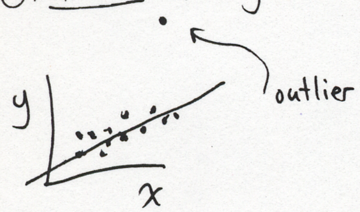
<u>Soln</u>: Use a regularizer term in the minimization

$$\hat{a} = \arg\min_a \sum_i \frac{\left[y_i - \sum_j a_j \phi_j(x_i)\right]^2}{2\sigma^2} + \lambda \sum_i a_i^2$$

- we penalize large coefficients a_i because they suggest we are fitting to noise/correlations
- We can set λ by cross validation; we leave out one data point $\{x_i, y_i\}$, fit to the rest, and predict the i^{th} point $\Longrightarrow$ goodness of prediction 

II EM

(A) Motivation: In regression setup, it'd be nice to throw away outliers & fit to "true" data



Imagine y_i's are generated by 2 processes

w/ prob π, $y = ax + b + \varepsilon$

$(1-\pi)$, $y \sim$ uniform noise

We can think of each data point (x_i, y_i) as coming w/ an extra label specifying which process generated it

$$X_i = (x_i, y_i, \delta_i)$$

Model parameters: a, b, π

1) Assume we have an estimate of our model $a^{(n)}\ b^{(n)}\ \pi^{(n)}$

We now want to estimate δ_i by $E[\delta_i \mid a^{(n)}, b^{(n)}, \pi^{(n)}]$

$$E[\delta_i \mid a^{(n)} b^{(n)} \pi^{(n)}] = 1\, P(\delta_i = 1 \mid a^{(n)} b^{(n)} \pi^{(n)}) + \underbrace{[0 \cdot P(\delta_i = 0 \mid a^{(n)} b^{(n)} \pi^{(n)})]}_{0}$$

$$\hookrightarrow P(\delta_i = 1 \mid a^{(n)}, b^{(n)}, \pi^{(n)}) = \frac{P(X_i = \text{observed value} \mid \delta_i = 1, a^{(n)}, b^{(n)}, \pi^{(n)})\, \underbrace{P(\delta_i = 1)}_{\pi^{(n)}}}{P(X_i = \text{observed value} \mid \delta_i = 1, \ldots)\, \underbrace{P(\delta_i = 1)}_{\pi^{(n)}} + P(X_i = \ldots \mid \delta_i = 0)\, \underbrace{P(\delta_i = 0)}_{(1-\pi^{(n)})}}$$

$$\hookrightarrow P(X_i = \text{observed value} \mid \delta_i = 1, a^{(n)}, b^{(n)}, \pi^{(n)}) \propto \exp\left(\frac{(y_i - ax - b)^2}{2\sigma^2}\right)$$

2) Given $E[\delta_i \mid \ldots]\ \forall i$, we can re-estimate $a^{(n+1)}, b^{(n+1)}, \pi^{(n+1)}$ with ~~[illegible]~~ weighted least squares

$$a^{(n+1)}, b^{(n+1)} = \arg\min_{a,b} \sum_i \frac{(y_i - ax - b)^2}{2\sigma^2} w_i \quad \text{where } w_i = E[\delta_i \mid \ldots]$$

We can set $\pi^{(n)} = \frac{1}{N} \sum_{i=1}^{N} w_i$

Li & Shum cont'd

	frame ← 49,800 →
model ↑ 246 ↓	0 0 0 0 1 0 0 0 0

hidden variable is binary vector with a '1' specifying which model this frame belongs to
equivalent to r_i from before

- each model, we see in 60-172 contiguous frame sequences in the data

Problems:

1) We don't have enough data to fit the dynamics for each of our 246 models
2) Can't compute E-step of EM

$$\mathbb{E}[r_i \mid \text{model}] = \ldots\ldots\ldots$$

- we first need to calculate the likelihood of observed frames
- $\underset{2}{\underline{\text{this}}}$ turns out to be intractable

i.e.

$$\begin{array}{lcccc} \text{model}(a) & X_1^{(a)} & \longrightarrow & X_2^{(a)} & \cdots \\ & \downarrow & & \downarrow & \\ & y_1^{(a)} & & y_2^{(a)} & \\ \text{model}(b) & X_1(b) & \longrightarrow & X_2(b) & \cdots \\ & \downarrow & & \downarrow & \\ & y_1(b) & & y_2(b) & \end{array}$$

average over all states $X_1^{(a)}$

$$P(y_1^{(a)} = \text{measured} \mid r_1 = a, \ldots) = \int P(y_1^{(a)} = \text{meas} \mid X_1^{(a)} \ldots) P(X_1^{(a)}) \, dX$$

$$P(y_2^{(a)} = \text{measured} \mid r_2 = a, \ldots) = \sum_{r_1} \underbrace{(\ldots)}_{r_1 = a} + \underbrace{(\ldots)}_{r_1 = b}$$

$$P(y_N^{(a)} = \text{measured} \mid r_n = a, \ldots) \sim \text{exponential in } N$$

for each previous frame, we need to consider both cases of model (a) and model (b)