4/13/18

Note Title 1/8/2000

Non-linear regression

- associated cost is nonlinear in its parameters
- Logistic regression

Logistic sigmoid function

18th century Verhulst

f population

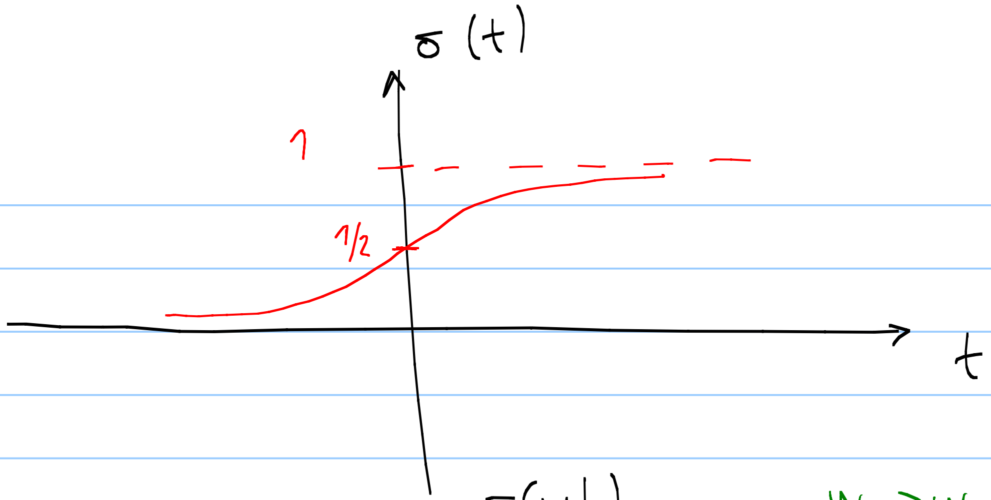
growth rate $\frac{df}{dt} = \underset{\uparrow \text{ current population}}{f}\,(\overset{\text{max capacity}}{1} - f)$ $\underbrace{}_{\text{remaining capacity}}$ $f(0) = 1/2$

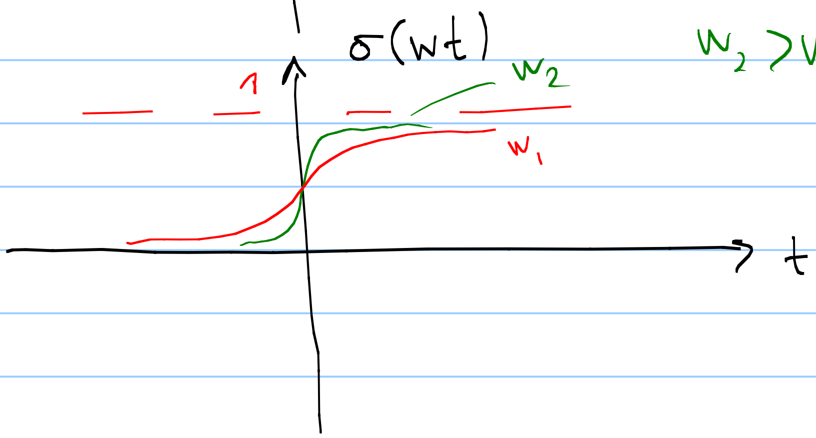
$$\boxed{\sigma(t) = \frac{1}{1+e^{-t}}}$$

$$\sigma'(t) = \frac{-(-e^{-t})}{(1+e^{-t})^2} = \underbrace{\frac{1}{1+e^{-t}}}_{\sigma(t)} \cdot \underbrace{\frac{e^{-t}}{1+e^{-t}}}_{1-\sigma(t)}$$

$$1 - \frac{1}{1+e^{-t}} = \frac{1+e^{-t}-1}{1+e^{-t}} = \frac{e^{-t}}{1+e^{-t}}$$

$$\sigma'(t) = \sigma(t)\,(1-\sigma(t))$$





$w_2 > w_1$

Logistic Regression

- a dataset $\{\bar{x}_p, y_p\}$, $p = 1, \ldots, P$ is distributed sigmoidally

$$\sigma(b + \bar{x}_p^T \bar{w}) \approx y_p, \quad p = 1, \ldots, P$$

- Least Squares approach

$$g(b, \bar{w}) = \sum_{p=1}^{P} \left(\sigma(b + \bar{x}_p^T \bar{w}) - y_p \right)^2$$

- Compact notation

$$\tilde{x}_p = \begin{bmatrix} 1 \\ \bar{x}_p \end{bmatrix}_{(N+1) \times 1} \qquad \tilde{w} = \begin{bmatrix} b \\ \bar{w} \end{bmatrix}_{(N+1) \times 1}$$

$$g(\tilde{w}) = \sum_{p=1}^{P} \left(\sigma(\tilde{x}_p^T \tilde{w}) - y_p \right)^2$$

$$\nabla_{\tilde{w}}\, g(\tilde{w}) = \sum_{p=1}^{P} \nabla \left(\sigma(\tilde{x}_p^T \tilde{w}) - y_p \right)^2$$

$$= \sum_{p=1}^{P} 2 \left(\sigma(\tilde{x}_p^T \tilde{w}) - y_p \right) \cdot \nabla_{\tilde{w}} \left(\sigma(\tilde{x}_p^T \tilde{w}) - y_p \right)$$

$$= \sum_{p} 2 \left(\sigma(\tilde{x}_p^T \tilde{w}) - y_p \right) \cdot \sigma(\tilde{x}_p^T \tilde{w}) \left(1 - \sigma(\tilde{x}_p^T \tilde{w}) \right) \cdot \underbrace{\nabla (\tilde{x}_p^T \tilde{w})}_{\tilde{x}_p}$$

Gradient descent

$$\tilde{w}_{k+1} = \tilde{w}_k - \alpha_k \cdot \nabla g(\tilde{w}_k)$$

Regularization

original cost $\qquad \min_{\tilde{w}} \; g(\tilde{w})$

regularization parameter

regularized cost $\qquad \min_{\tilde{w}} \left(g(\tilde{w}) + \lambda R(\tilde{w}) \right)$

Examples of $R(\tilde{w})$

- $R(\tilde{w}) = \|C \tilde{w}\|_2^2$ $\qquad$ incorporation of prior knowledge
- $R(\tilde{w}) = \|\tilde{w}\|_1^2$ $\qquad$ ℓ_1 norm enforces sparcity
- avoid overfitting $\qquad R(\tilde{w}) = \|\tilde{w}\|_2^2$
- convexify $g(\tilde{w})$ $\qquad R(\tilde{w}) = \|\tilde{w}\|_2^2$