④ Supervised learning I

* Supervised : $D = \{x^i, y^i\}$ (features, outputs) * Example { y : temperature; x_1 : day; x_2 : place;

a) Linear regression and Polynomic regression

* { $\min_\theta \; f(\theta) = MSE(\theta) = \| y - X\theta \|_2^2$; The output $y \in \mathbb{R}^N$

  where $X = [1, x, x^2, \dots]$ for one feature and $X = [1, x_1, x_2, \dots, x_1^2, x_1 x_2, \dots]$

* { We can use different norms / or penalty functions

* Regularization

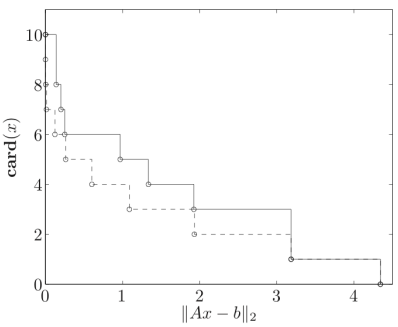
* L^2 regularization (Tikhonov regularization / Ridge regression)

$$\min \| y - X\theta \|_2^2 + \lambda \underbrace{\|\theta\|^2}_{\theta^T\theta}$$

- $\lambda > 0$
- Prior knowledge of θ
- Penalize large values of θ
- Also useful when X is ignore but bad conditioned ; Elasticity ?

$$(\underbrace{X^TX + \lambda I}_{\text{full rank}})\theta = X^Ty$$

* Other possible regularization
  * L^1 regularization $|\theta|$ (Lasso problem)



$$\begin{cases} \min \| y - X\theta \|_2^2 \\ \text{s.t} \quad \text{card}(\theta) \leq k \end{cases}$$

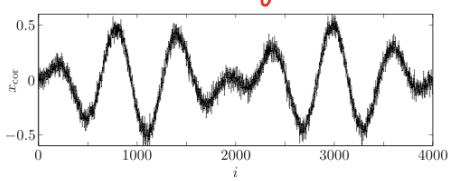
asking for a sparse representation of θ

good approximation

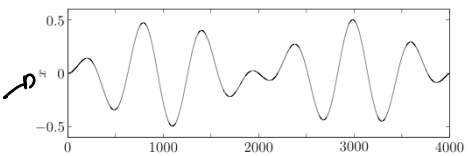
$$\min \| y - X\theta \|_2^2 + \lambda \|\theta\|_1$$

* $\|\nabla\theta\|^2$ → smoothing (Laplacian)

$$\min \|y - I\theta\|_2^2 + \lambda\|\nabla\theta\|^2$$



y



θ

In variational form: $\min_{\theta} \int(y-\theta)^2 + \varepsilon\int\nabla\theta^2 = J(\theta)$

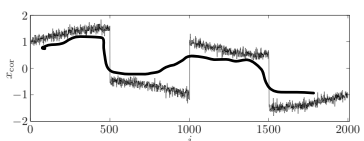
Gateaux derivative: $\frac{\partial J(\theta)}{\partial\theta}(v) = 2\int(y-\theta)v + \varepsilon\int\nabla\theta\nabla v = 0 \quad \forall v$

$$\int v\theta + \varepsilon^2\int\nabla v\nabla\theta = \int yv \quad \forall v \Rightarrow \text{Strong form} \begin{cases} \varepsilon^2\Delta\theta + \theta = y \end{cases}$$

$$[I + \varepsilon^2 K]\theta = \tilde{y}$$

* $|\nabla\theta|$ → Total variation

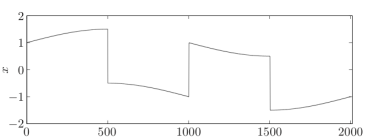
$$\min \|y - I\theta\|_2^2 + \lambda|\nabla\theta|$$



y

$\theta^T\theta$

Denoising

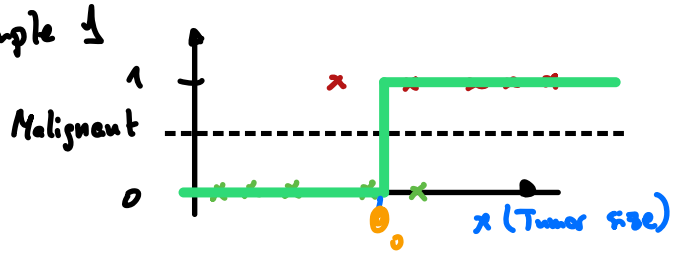


θ

b) Logistic regression / classification

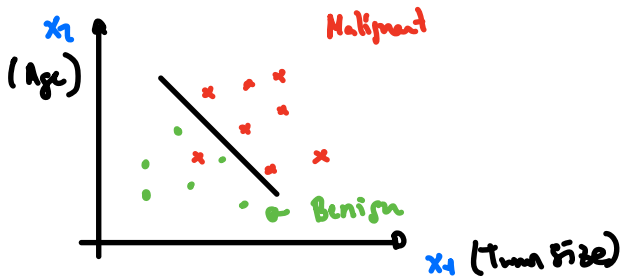
* Classification: the output variable $y \in \{0, 1\}$

* Example 1



θ_0

* Example 2



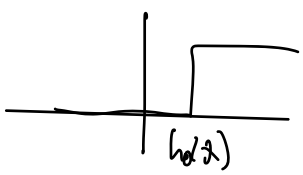
* Hypothesis space / hypothesis function

$$h_\theta(x) = X\theta$$

* Thresholding (Heaviside function)

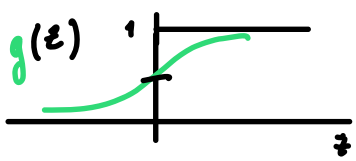
$$g(h_\theta(x)) = \begin{cases} 0 & h_\theta(x) < 0 \\ 1 & h_\theta(x) \geq 0 \end{cases}$$

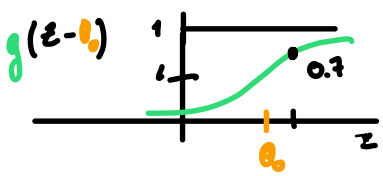
$g(h_\theta(x))$ thresholding



* Signoid function: (Logistic function)

$$g(z) = \frac{1}{1+e^{-z}}$$

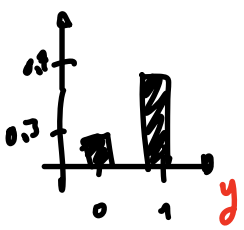




* Interpretation of $g(h_\theta(x))$:

Estimated probability that $y=1$ on input x

For some x_1 so $g(h_\theta(x_1)) = 0.7 \Rightarrow$ 70% chance of human being malignant



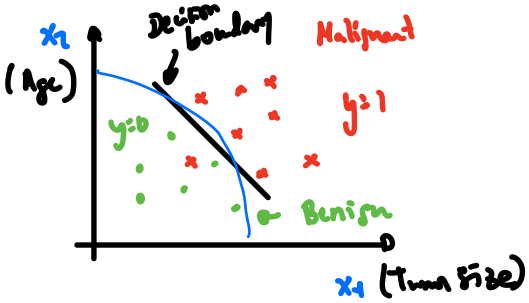
$$g(h_\theta(x)) = p(y=1 \mid x=x; \theta)$$

$$1 - g(h_\theta(x)) = 1 - p(y=1 \mid x=x; \theta) = p(y=0 \mid x=x; \theta) = 0.3$$

* Decision boundary

$$g(\theta^T x) = \begin{cases} 1 & \theta^T x \geq 0 \\ 0 & \theta^T x < 0 \end{cases}$$

θ linear



* Cost function: Negative loglikelihood / cross-entropy

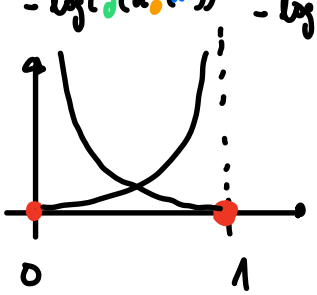
$$J_\theta(\theta) = \frac{1}{m} \sum_{i=1}^{m} (1-y^i)(-\log(1-g(h_\theta(x^{(i)})))) + y^i(-\log(g(h_\theta(x^{(i)}))))$$

$y_i = 0 \Rightarrow g = 1 \Rightarrow -\log(1-g) \to \infty$

$(y_i = 0)$

$-\log(g(h_\theta(x^{(i)})))$

$-\log(1-g(h_\theta(x^{(i)})))$ $(y_i = 1)$



$$\begin{cases} \text{if } y^i = 0 \Rightarrow \text{penalize } 1 = g(h_\theta(x^{(i)})) \\ \text{if } y^i = 1 \Rightarrow \text{penalize } 0 = g(h_\theta(x^{(i)})) \end{cases}$$

* Property of Sigmoid function

$$g(z) = \frac{1}{1+e^{-z}} \; ; \; g'(z) = g(z)(1-g(z))$$

* Gradient

$\frac{\partial h_\theta}{\partial \theta} \sim$ $\quad h = \theta x$

$$\frac{\partial J_\theta(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(1-y^i)\frac{-1}{(1-g(h_\theta(x^{(i)})))}\,-(g(h_\theta(x^{(i)}))(1-g(h_\theta(x^{(i)}))\,x^{(i)}$$

$$+\;(y^i)\frac{-1}{(g(h_\theta(x^{(i)})))}\,(g(h_\theta(x^{(i)}))(1-g(h_\theta(x^{(i)}))\,x^{(i)}$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[(1-y^i)(g(h_\theta(x^{(i)}))-(y^i)(1-g(h_\theta(x^{(i)}))\right]x^{(i)}$$

the gradient is similar to polynomic regression

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[g(h_\theta(x^{(i)}))-y^i\right]x^{(i)} = 0$$

$$\min_\theta \tfrac{1}{2}\|X\theta - y\|^2 = J_\theta(\theta) \qquad \vec{X}^T X\theta = \vec{X}^T y$$

$$\frac{\partial J_\theta(\vec{\theta})}{\partial \theta} = X^T X\theta - X^T y = 0 \qquad \nabla J(\theta^*) = 0$$

$$= X^T[X\theta - y]$$

* Remarks

- * You can use gradient method: $\theta_{k+1} = \theta_k - \tau \nabla J_\theta(\theta_k)$
  - * Stochastic gradient, randomly select l values of the data

$$\nabla J_\theta(\theta_k) \simeq \frac{1}{m}\sum_{i=1}^{l}\left[g(h_\theta(x^{(i)}))-y^i\right]x^{(i)}$$

  - * Regularization also important

$$\min_\theta \frac{1}{m}\sum_{i=1}^{m}(1-y^i)(-\log(1-g(h_\theta(x^{(i)}))) + y^i(-\log(g(h_\theta(x^{(i)}))))$$

$$+\lambda\,\theta^T\theta$$

Typically we take hundreds of data to compute the stochastic gradient when we have used all the data we say that we spend an epoch

For L_1 norm regularization use: Proximal methods

$$\frac{\partial J_\theta(\theta)}{\partial \theta} = \frac{1}{m}\sum_{i=1}^{m}\left[g(h_\theta(x^{(i)}))-y^i\right]x^{(i)} + \lambda\,\theta$$

- * Gradient expensive → Stochastic
- * Stopping criteria is expensive → $\|\nabla J\|$, ..., n^o iterations
- * Computing cost expensive → We don't compute it
  - to τ → Exact line-search / Backtracking → τ very small but fixed