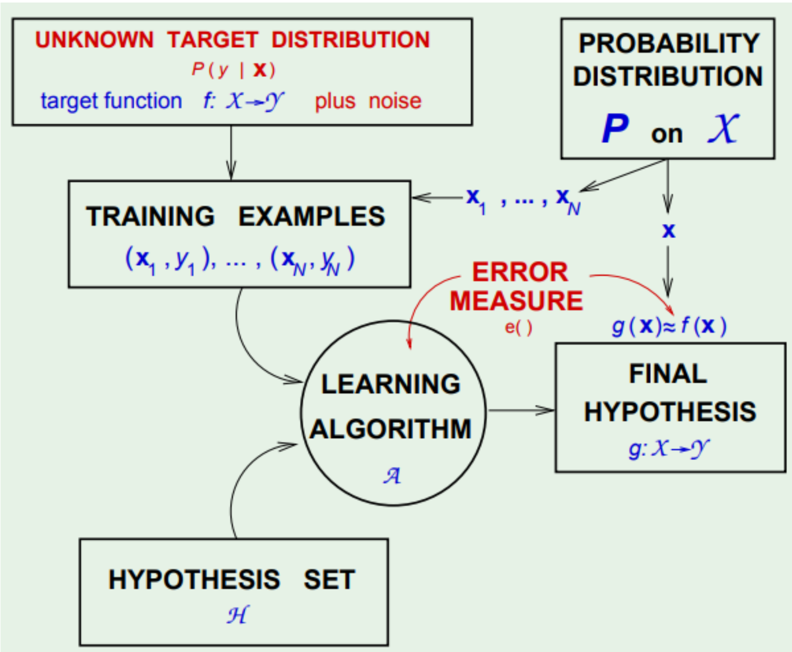
EDX ML

Yaser Abu-Mostafa

10/22/2017

https://courses.edx.org/courses/course-v1:CaltechX+CS1156x+3T2017/course/

1 Week #1



Good generalization: $E_{in}(g) \approx E_{out}(g)$

Learning: $g \approx f \iff E_{out}(g) \approx 0$

...

2 Week #2

$$X = \begin{bmatrix} \text{——} & \mathbf{x}_1^T & \text{——} \\ \text{——} & \mathbf{x}_2^T & \text{——} \\ & \vdots & \\ \text{——} & \mathbf{x}_N^T & \text{——} \end{bmatrix} \text{ — input data matrix, } \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \text{ — target vector}$$

$$E_{in}(\mathbf{w}) = \frac{1}{N} \|X\mathbf{w} - \mathbf{y}\|^2$$

$$\nabla E_{in}(\mathbf{w}) = \frac{2}{N} X^T (X\mathbf{w} - \mathbf{y}) = \mathbf{0}$$

$$\mathbf{w} = X^\dagger \mathbf{y} \quad \text{where} \quad X^\dagger = (X^T X)^{-1} X^T$$

$X^\dagger$ — pseudo-inverse, `numpy.linalg.pinv(X)`

2.1 Error Measure

$$E_{in} = \frac{1}{N} \sum_{n=1}^{N} e\left(h(x_n), f(x_n)\right)$$

$$E_{out} = \mathbb{E}_{\mathbf{x}} \left[e\left(h(\mathbf{x}), f(\mathbf{x})\right)\right]$$

2.2 Noisy Targets

Instead of $y = f(\mathbf{x})$ we use target *distribution*: $P(y|\mathbf{x})$

$(\mathbf{x}, y)$ is now generated by joint distribution: $P(x)P(y|\mathbf{x})$

Noisy target $\equiv$ deterministic target $f(x) = \mathbb{E}(y|\mathbf{x})$ $\pm$ some noise $(y - f(\mathbf{x}))$

3 Week 3

3.1 Theory of generalization

Hoeffding's inequality:

$$P\left[|E_{in}(g) - E_{out}(g)| > \epsilon\right] \leqslant 2Me^{-2\epsilon^2 N}$$

This inequality is a form of the large numbers law. The statement $E_{in} = E_{out}$ is *PAC* (probably approximately correct).

M – number of functions in the hypothesis set $\mathcal{H}$, N – number of *dichotomy* points

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \ldots, \mathbf{x}_N)| \quad \text{— growth function}$$

When you get all possible hypotheses, all possible dichotomies, you say that the hypothesis set $\mathcal{H}$ *shattered* the points – broke them in all possible 2^N ways.

k — *break point*, minimum N, where number of dichotomies cannot reach 2^N

If $k = \infty$ (no break point), then $m_{\mathcal{H}}(N) = 2^N$

$B(N,k)$ — maximum number of dichotomies on N points, with break point k

$$B(N,k) \leq \alpha + 2\beta = (\alpha + \beta) + \beta = B(N-1,k) + B(N-1,k-1) \quad \ldots$$

$$B(N,k) = \sum_{i=0}^{k-1} \binom{N}{i}$$

$$\text{Note:} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{1}{k!} \underbrace{n(n-1)\cdots(n-k+1)}_{k \text{ times}} \sim n^k$$

$$m_{\mathcal{H}}(N) \leqslant B(N,k) = \sum_{i=0}^{k-1} \binom{N}{i} \sim N^{k-1} \quad \text{— polynomial}$$

Note: order of the polynomial depends on the break point.

The Vapnik-Chernovenskis inequality:

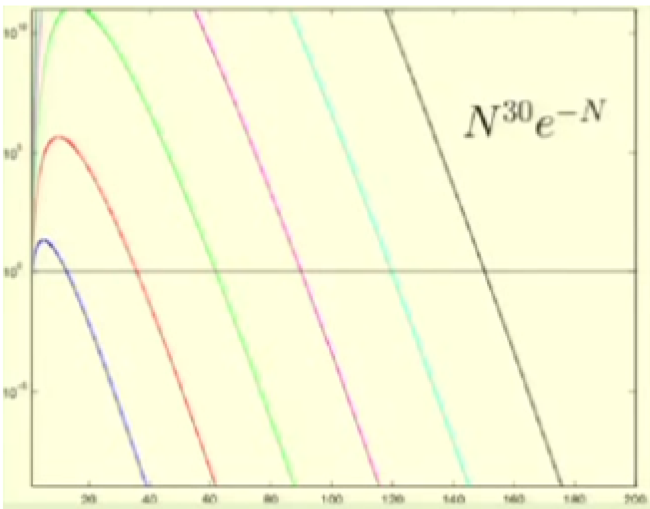
$$P\big[\,|E_{in}(g) - E_{out}(g)| > \varepsilon\,\big] \leqslant 4m_{\mathcal{H}}(2N)\,e^{-\frac{1}{8}\varepsilon^2 N} \quad (\triangleq \delta)$$

4 Week 4

4.1 VC dimensions

The VC dimension $d_{VC}(\mathcal{H})$ is the the most points $\mathcal{H}$ can shatter.

$$m_{\mathcal{H}}(N) \leq \sum_{i=0}^{d_{VC}} \binom{N}{i} \sim N^{d_{VC}}$$



$$\text{Rule of thumb:} \quad P \lesssim 10^{-1} \Longleftrightarrow N \gtrsim 10 d_{VC}$$

$$\text{(probability bound)}\ \delta = 4\,m_{\mathcal{H}}(2N)\,e^{-\frac{1}{8}\epsilon^2 N} \Longleftrightarrow \epsilon = \sqrt{\frac{8}{N} \ln \frac{4\,m_{\mathcal{H}}(2N)}{\delta}} \quad (\triangleq \Omega)$$

With probability $P = 1 - \delta$, $|E_{out} - E_{in}| \leq \Omega(N,\mathcal{H},\delta)$

Generalization bound:

$$E_{out} \le E_{in} + \Omega$$

$$|\mathcal{H}|\uparrow \Rightarrow E_{in}\downarrow \quad \text{but} \quad \Omega\uparrow$$

4.2 Bias-Variance Tradeoff

$$E_{out}(\mathcal{D}) = \mathbb{E}_{\mathbf{x}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

Average hypothesis:

$$\bar{g}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}\left[g^{(\mathcal{D})}(\mathbf{x})\right]$$

$$\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x})\right)^2\right] = \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]}_{\mathbf{var}(\mathbf{x})} + \underbrace{\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2}_{\mathbf{bias}(\mathbf{x})}$$

$$\mathbf{bias} = \mathbb{E}_{\mathbf{x}}\left[\left(\bar{g}(\mathbf{x}) - f(\mathbf{x})\right)^2\right]$$

$$\mathbf{var} = \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}\left[\left(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x})\right)^2\right]\right]$$