* În orașul B prob. a priori de a avea loc o crimă este invers proporțională cu nr. de bărbați adulți din orașul B. Într-un proces care este probabilitatea ca acuzatul să fie găsit vinovat știind că la locul crimei a fost găsită o mostră de ADN care coincide cu ADN-ul acuzatului. Probabilitatea ca două ADN-uri să coincidă este de 10^{-6} iar în oraș sunt $5 \cdot 10^4$ bărbați adulți.

Sol.

G = "ev. că acuzatul este vinovat"

E = "ev. ca ADN-ul acuzatului să coincidă cu mostra prelevată"

$\Rightarrow P(G) = \frac{1}{5 \cdot 10^4}$ și $P(E/\overline{G}) = \frac{1}{10^6}$

- $P(E/G) = 1$ căci dacă acuzatul este criminalul atunci cele 2 ADN-uri coincid.
- Sist complet de ev. $\{G, \overline{G}\}$ $\Longrightarrow P(\overline{G}) = 1 - P(G)$
- Th Bayes $P(G/E) = \dfrac{P(G)P(E/G)}{P(G)P(E/G) + P(\overline{G})P(E/\overline{G})}$

$= 0{,}952381859 41$

....

* Într-o urnă se află n bile. Sunt posibile $(n+1)$ ipoteze asupra numărului de bile albe din urnă. Se notează cu H_i ipoteza că în urnă sunt i bile albe. Presupunem că toate ipotezele sunt egal probabile. Din urnă se extrage la întâmplare o bilă care este albă. Care este ipoteza cu cea mai mare probabilitate a posteriori? Justificați.

Sol

- Sist. complet de evenimente $\{H_i\}_{i=0}^{n}$
- $P(H_i) = \frac{1}{n+1}$, $i = \overline{0,n}$
- B = "ev. extragerii unei bile albe"
- (form. lui Bayes $P(H_i|B) = \dfrac{P(H_i)P(B|H_i)}{\sum\limits_{j=0}^{n} P(H_j)P(B|H_j)}$
- $\Rightarrow P(H_i|B) = \dfrac{\frac{1}{n+1}\cdot\frac{i}{n}}{\frac{1}{n(n+1)}\cdot\sum\limits_{j=0}^{n} j} = \dfrac{2i}{n(n+1)}$
- $\max\limits_{i} P(H_i|B) = \dfrac{2n}{n(n+1)} = \dfrac{2}{n+1}$ și corespunde ip. H_n

* (Semnale în zgomotul gaussian) Fie s_0 și $s_1 \in \mathbb{R}$ date și $Z \sim N(0, \sigma^2)$ cu σ^2 dată. Fie $P[Y=1] = 1/2$ și fie $X = \begin{cases} s_0 & \text{dacă } Y=0 \\ s_1 & \text{dacă } Y=1 \end{cases}$. Construiți g^* funcție de decizie bayesiană și calculați L^* eroarea sa de misclasare pt $s_0 = -1$, $s_1 = 1$ și $\sigma^2 = 1$.

<u>Sol.</u>

$$g^* : \mathbb{R} \to \{0, 1\} \quad \text{cu} \quad g^*(x) = \begin{cases} 1 & \text{dacă } \eta(x) > 1/2 \\ 0 & \text{în rest} \end{cases}$$

unde $\eta(x) = \dfrac{p f_1(x)}{p f_1(x) + (1-p) f_0(x)}$, $p = P[Y=1]$, f_0 și f_1 d.p. a $X|Y=0$ respectiv $X|Y=1$

$$\Rightarrow g^*(x) = \begin{cases} 1 & \text{dacă } p f_1(x) > (1-p) f_0(x) \\ 0 & \text{în rest} \end{cases}$$

$$\Rightarrow g^*(x) = \begin{cases} 1 & \text{dacă } 2(s_1 - s_0)x + s_0^2 - s_1^2 > 0 \\ 0 & \text{în rest} \end{cases} \Rightarrow g^*(x) = \begin{cases} 1 & \text{dacă } x > \frac{s_1 + s_0}{2} \\ 0 & \text{în rest} \end{cases}$$

$$L^* \overset{\text{def}}{=} \int_{\mathbb{R}} \min(\eta(x), 1-\eta(x)) f_X(x)\, dx = \frac{1}{2} - \frac{1}{4} \int_{\mathbb{R}} |f_1(x) - f_0(x)|\, dx$$

$$f_X(x) = p f_1(x) + (1-p) f_0(x)$$

cu val. integralei luată din Tabelul lui $N(0,1)$

$$= \frac{1}{2} - \int_0^1 f(x)\, dx = 0{,}158655$$

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Sem 2

Haykin - Neural Networks and Learning Machines 3.1 (pag 69 - 73)

w_{MAP} (maximum aposteriori) estimator $= w_{RMAL}$ (~~regularized least squares estimator~~) (regularized maximum-likelihood) estimate

în modelul de regresie liniară (pornind de la slide-ul de lecture 3b) în ipoteza zgomotului gaussian

SSL

Linear regression model: preliminary consideration

- an unknown stochastic environment produce un regresor $\underline{X} \in \mathcal{M}_{p\times 1}(\mathbb{R})$
- the response of the supervisor is $y \in \mathbb{R}$
- the LM propose a linear regression model, i.e

$$y = \sum_{j=1}^{p} w_j x_j + \varepsilon$$

unde $w_1, \dots, w_p$ ~~sunt fi~~ denote a set of fixed, but unknown parameters meaning that the environment is stationary

The additive term ε, representing the expectational error of the model, accounts for our ignorance about the environment.

$$\underset{1\times 1}{y} = \underset{1\times p}{\underline{w}^T} \cdot \underset{p\times 1}{\underline{x}} + \varepsilon \qquad \text{cu } \underline{w} \in \mathcal{M}_{p\times 1}(\mathbb{R})$$

$$\underset{m\times 1}{Y} = \underset{m\times p}{X^T}\, \underset{p\times 1}{\underline{w}} + \underset{m\times 1}{\varepsilon}$$

- The loss functions:
  - Sum of squared errors

$$\mathcal{E}(\underline{w}) = \frac{1}{2}\sum_{i=1}^{n} \underbrace{\left(y_i - \underline{x}_i^T \underline{w}\right)^2}_{\ell(y,\, h(\underline{x}))} \qquad \underline{x}^T\underline{w} = h$$

$$\Longrightarrow \hat{w}_{ML} = \min_{\underline{w}} \mathcal{E}(\underline{w}) = \min_{\underline{w}} (Y - X^T\underline{w})^T \cdot (Y - X^T\underline{w}) = \min_{\underline{w}} \|Y - X^T\underline{w}\|^2 = \min \|\varepsilon\|^2$$

  - Regularized least squared errors

$$\hat{w}_{RMAL} = \min_{\underline{w}} \left[\mathcal{E}(\underline{w}) + \frac{\lambda}{2}\|w\|^2\right] = \min_{\underline{w}} \left[\mathcal{E}(\underline{w}) + \frac{\lambda}{2}\underline{w}^T\underline{w}\right]$$

↑ norma euclidiană

statistică: ridge regression

Neural Nets: weights decay

1/4

- Maximum a posteriori estimator

Assumption 1: Statistical Independence and Identical Distribution (i.i.d)
~~Assumption 2: Stationarity~~ The n examples, constituting the training sample, are statistically independent and identically distributed

Assumption 2: The $D_{\cancel{x}}$ is gaussian, i.e. the environment, responsible for generation of the training sample $\mathcal{T}$, is Gaussian distributed. More specifically

$$\varepsilon_i \sim N(0, \sigma^2) \qquad p(\varepsilon_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right), \; i=1,..,n$$

environment

Assumption 3 The environment is stationary, which means that the parameters vector $\underline{w}$ is fixed, but unknown, throughout the n trials of the experiment.

$$\Rightarrow \quad E[y_i] = \underline{x}_i^T \underline{w} \quad , \overline{i=1,..,n}$$

$$var[y_i] = \cancel{E}[(y_i - E[y_i])^2] = var[\underline{x}_i^T \underline{w} + \varepsilon_i] = var[\varepsilon_i] = \sigma^2$$

Conform th. Bayes pt. densităţi

$$\Rightarrow \quad p(\underline{w} \mid \cancel{Y}, X) = \frac{p(\underline{w})\, p(\cancel{Y} \mid \underline{w}, \underline{X})}{p(\cancel{Y})}$$

$$\propto p(\underline{w})\, \underbrace{p(Y \mid \underline{w}, X)}_{\text{likelihood function}}$$

$$\cancel{\hat{w}_{MAP} = \max_{w} [p(\underline{w})\, p(Y \mid \underline{w}, X)]} \quad \cancel{\ell(y, h(x))}$$

$$\cancel{\text{Dar } p(Y \mid \underline{w}, X) =}$$

$$w_{MAP} = \arg\max_{\underline{w}} \underbrace{\ln(p(\underline{w} \mid \cancel{Y}, \underline{X}))}_{\ell(\cancel{y}, h(\underline{x}))}$$

Din Ass 2 $\Rightarrow p(\underline{w} \mid y_i, \underline{x}_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_i - \underline{x}_i^T \underline{w})^2\right)$, $i=1,..,n$

$\underline{x}_i^T \underline{w}$

Ass 1

$$\Rightarrow \underbrace{p(\underline{w} \mid \underline{Y}, X)}_{\text{likelihood function}} = \prod_{i=1}^{n} p(\underline{w} \mid y_i, \underline{x}_i) = \frac{1}{(2\pi\sigma^2)^{n/2}} \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma^2}(y_i - \underline{x}_i^T \underline{w})^2\right)$$

$$p(\underline{w}|\underline{Y},X) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \underline{x}_i^T\underline{w})^2\right)$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big[-\frac{1}{2\sigma^2}\underbrace{(Y-X^T\underline{w})^T(Y-X^T\underline{w})}_{\mathcal{E}(\underline{w})}\Big]$$

$$\Rightarrow \ln p(\underline{w}|\underline{Y},X) = -\frac{1}{2\sigma^2}\mathcal{E}(\underline{w}) - \frac{n}{2}\ln(2\pi\sigma^2)$$

The only other source of information that remains to be accounted for is that contained in the prior $p(\underline{w})$. ~~Looking the~~ More specifically, the p elements of the weight vector $\underline{w}$ are themselves assumed to be i.i.d. gaussian elements, i.e.

$$p(w_j) = \frac{1}{\sqrt{2\pi\sigma_w^2}} \exp\left(-\frac{w_j^2}{2\sigma_w^2}\right) \quad j=1,..,p$$

$$\Rightarrow \quad p(\underline{w}) = \prod_{j=1}^{p} p(w_j) = \frac{1}{(2\pi\sigma_w^2)^{p/2}} \exp\left[-\frac{1}{2\sigma_w^2}\sum_{j=1}^{p} w_j^2\right]$$

$$= \frac{1}{(2\pi\sigma_w^2)^{p/2}} \exp\Big[-\frac{1}{2\sigma_w^2}\underbrace{w^T w}_{\|w\|^2}\Big]$$

euclidean norm

$$\Rightarrow \ln p(\underline{w}) = -\frac{1}{2\sigma_w^2}\|w\|^2 - \frac{p}{2}\ln(2\pi\sigma_w^2)$$

Given $p(\underline{w}|Y,X) \propto p(\underline{w})\,p(Y|\underline{w},X) \Rightarrow$

$$\ln p(\underline{w}|Y,X) \propto \ln p(\underline{w}) + \ln p(Y|\underline{w},X)$$ in ln coordinates

$$\Rightarrow \hat{w}_{MAP} = \arg\max_{\underline{w}}\left[-\frac{1}{2\sigma^2}\mathcal{E}(\underline{w}) - \frac{1}{2\sigma_w^2}\|\underline{w}\|^2\right] = \min_{\underline{w}}\Big[\mathcal{E}(\underline{w}) + \frac{1}{2}\underbrace{\left(\frac{\sigma}{w_w}\right)^2}_{\lambda}\|w\|^2\Big]$$

$\longrightarrow$

Obs

1) The Bayesian paradigm for parameter estimation, rooted in the Bayes Theorem and exemplified by the MAP estimator exploits all the conceivable information about the parameter vector w. In contrast, the ML estimator lies only on the likelihood function ignoring the prior. $\Longrightarrow$ overfitting and poor generalization

2.) The ML estimator relies solely on the observation model $(y, \underline{X})$ and may therefore lead to a nonunique solution. To enforce uniqueness and stability on the solution, the prior $p(\underline{w})$ has to be incorporated into the formulation of the estimator; this is done in the RML estimator or automatically in the MAP estimator.

* Găsiți formula lui $\underline{w}^* \in \mathbb{R}^{n+1}$, ponderile unui perceptron ADALINE, ce minimizează, pe o mulțime de antrenare dată, SSE regularizată cu termenul $\frac{\lambda}{2}\|\underline{w}\|^2$ cu $\lambda > 0$ dat.

Soluție

- Fie $\{(\underline{x}_i, t_i)\}_{i=\overline{1,m}}$, $\underline{x}_i \in \mathbb{R}^n$, $t_i \in \mathbb{R}$, mulțimea de antrenare
- Fie $y(\underline{x}) = \tilde{\underline{w}}^T \cdot \tilde{\underline{x}}$ ieșirea perceptronului ADALINE

  unde $\tilde{\underline{x}} = [1, \underline{x}]^T$ și $\tilde{\underline{w}} = [w_0, w_1, \dots, w_n]^T$.
- Fie $\underline{X} = [\tilde{x}_1, \dots, \tilde{x}_m]$ matricea $(n+1) \times m$ a exemplelor

  și $\underline{t} = [t_1, \dots, t_m]^T$ vectorul $m \times 1$ al etichetelor
- Fie de cost în paradigma SSE regularizată este:

$$J(\tilde{\underline{w}}) = \frac{1}{2}\sum_{i=1}^{m}(t_i - y(\tilde{\underline{x}}_i))^2 + \frac{\lambda}{2}\|\tilde{\underline{w}}\|^2$$

(Pt. simplificare, în cele ce urmează vom renunța la "~" și la "_")

Se observă că

$$J(w) = \frac{1}{2}(t - X^T w)^T (t - X^T w) + \frac{\lambda}{2} w^T \cdot w$$

$$\Rightarrow \nabla J(w) = -X(t - X^T w) + \lambda w$$

$$\Rightarrow \nabla J(w) = 0 \Rightarrow (XX^T) w + \lambda w = X \cdot t$$

$$[(XX^T) + \lambda I_{n+1}] w = Xt$$

$$\Rightarrow w^* = [(XX^T) + \lambda I_{n+1}]^{-1} \cdot X \cdot t$$

Exercițiu. Sunteți concurent într-un joc spectacol în care un premiu este ascuns în spatele uneia dintre cele trei cortine. Veți câștiga premiul dacă alegeți ~~varianta~~ cortina corectă. După ce ați ales cortina, dar înainte ca ea să fie ridicată, prezentatorul ridică una dintre celelalte 2 cortine, arătând o scenă goală și vă întreabă dacă doriți să vă schimbați opțiunea curentă pentru cortina rămasă. Care este strategia optimă: păstrarea sau schimbarea opțiunii curente?

uşa = cortina

Rezolvare: $\Omega = \{a_1, a_2, a_3\} \times \{d_1, d_2, d_3\} \times \{m_1, m_2, m_3\}$.

(alegem ușa a_i; deschide ușa; mașina)

$|\Omega| = 12$ $\Omega = \{(a_i, d_j, m_k) \mid i \neq j, j \neq k\}$.

$= \{(a_1, d_2, m_1), (a_1, d_2, m_3), (a_1, d_3, m_1), \dots$

$\dots\}$.

Notăm cu A_i evenimentul să aleg ușa i

D_i evenimentul ca prezentatorul să deschidă ușa i

M_i evenimentul ca mașina să fie în spatele ușii = cortinei i.

Fără a pierde generalitatea, analizăm cazul în care aleg ușa = cortina = 1 și deschid ușa = cortina = 2.

Prezentatorul știe A_i și m_i.

$P(M_2 \mid A_1 \cap D_2) = 0.$ $P(A_1 \cap D_2) = P(A_1)\, P(D_2 / A_1) = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$

$P(M_1 \mid A_1 \cap D_2) = \frac{P(M_1 \cap A_1 \cap D_2)}{P(A_1 \cap D_2)} = \frac{P(D_2 \mid M_1 \cap A_1) \cdot P(M_1 \cap A_1)}{P(A_1 \cap D_2)} = \frac{\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{3}}{\frac{1}{6}} = \frac{1}{3}$

$P(M_3 \mid A_1 \cap D_2) = \frac{P(D_2 \mid M_3 \cap A_1)\, P(M_3 \cap A_1)}{\frac{1}{6}} = \frac{1 \cdot \frac{1}{3} \cdot \frac{1}{3}}{1/6} = \frac{2}{3}$