Formulario Machine Learning

Supervised Learning $D \subset \{(x,y) \mid x \in X, y \in Y\}$

- Classification $f: X \to Y$ with $X \subset \mathbb{R}^d$ and $Y = \{C_1, \dots, C_k\}$
- Regression $f: X \to Y$ with $X \subset \mathbb{R}^d$ and $Y = \mathbb{R}$

Unsupervised Learning $D \subset \{x \mid x \in X\}$

Reinforcement Learning $D = \{(a_i^1 \dots a_i^n), r_i \mid i \in 1 \dots |S|\}$

Classification evaluation

True error $error_D(h) = \Pr_{x \in D}[f(x) \neq h(x)]$ D is the probability distribution over X

Sample error $error_S(h) = \frac{1}{n} \sum_{x \in S} \delta(f(x) \neq h(x))$ S is the dataset

$\delta = \begin{cases} 1 & \text{if } f(x) \neq h(x) \\ 0 & \text{otherwise} \end{cases}$

True error interval $error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1-error_S(h))}{n}}$

Recall = $TP/(TP+FN)$ ability to avoid false negatives

Precision = $TP/(TP+FP)$ ability to avoid false positives

F_1-Score = $2(Precision \cdot Recall)/(Precision + Recall)$

Probability

Conditional probability: $P(a|b) = P(a \wedge b)/P(b)$ if $P(b) \neq 0$

Product rule: $P(a \wedge b) = P(a|b)P(b) = P(b|a)P(a)$

Sum rule: $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

Total probability: $P(a) = P(a|b)P(b) + P(a|\neg b)P(\neg b)$

Independence: X is independent from Y given Z if $P(x, y|z) = P(x|z)$

Bayes rule: $P(a|b) = P(b|a)P(a)/P(b)$

Bayesian Learning

Maximum a posteriori hypothesis: $h_{MAP} \equiv \underset{h \in H}{\arg\max}\, P(h|D) = \underset{h \in H}{\arg\max}\, P(D|h)P(h)$

Maximum likelihood hypothesis: $h_{ML} \equiv \underset{h \in H}{\arg\max}\, P(D|h)$

Probabilistic models for classification

Generative model: $w^T x + w_0 \to w = \Sigma^{-1}(\mu_1 - \mu_2)$; $w_0 = \frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2$

$P(C_1|x) = \sigma(w^T x + w_0)$

Likelihood: $\prod_{n=1}^{N} \left[\pi \mathcal{N}(x_n; \mu_1, \Sigma)\right]^{t_n} \left[(1-\pi)\mathcal{N}(x_n; \mu_2, \Sigma)\right]^{(1-t_n)}$

$= p(t \mid \pi, \mu_1, \mu_2, \Sigma, D)$

- Likelihood to be max: $\pi, \mu_1, \mu_2, \Sigma = \underset{\pi, \mu_1, \mu_2, \Sigma}{\arg\max} \log p(t \mid \pi, \mu_1, \mu_2, \Sigma, D)$

Optimal parameters

$$\pi = \frac{N_1}{N} \; ; \; \mu_1 = \frac{1}{N_1} \sum_{n=1}^{N} t_n x_n \; ; \; \mu_2 = \frac{1}{N_2} \sum_{n=1}^{N} (1-t_n) x_n$$

$$\Sigma = \frac{N_1}{N} S_1 + \frac{N_2}{N} S_2 \text{, with } S_i = \frac{1}{N_i} \sum_{n \in C_i} (x_n - \mu_i)(x_n - \mu_i)^T \text{, } i = 1, 2$$

Linear models for Classification

Least squares

Error function: $E(\tilde{W}) = \frac{1}{2} Tr\left\{(\tilde{X}\tilde{W} - T)^T (\tilde{X}\tilde{W} - T)\right\}$ to be minimize

- Solution: $\tilde{W} = \tilde{X}^{\dagger} T$

Model: $y(x) = \tilde{W}^T \tilde{x} = T^T (\tilde{X}^{\dagger})^T \tilde{x}$

Perceptron

Error function: $E(w) = \frac{1}{2} \sum_{n=1}^{N} (t_n - w^T x_n)^2$ to be minimize

Parameters update: $w_i \leftarrow w_i + \Delta w_i$

$\Delta w_i = -\eta \frac{dE}{dw_i} = \eta \sum_{n=1}^{N} (t_n - sign(w^T x_n))(x_{i,n})$ sign viene applicata solo alla fine!

SVM

- Problem: $w^*, w_0^* = \underset{w, w_0}{\arg\max} \frac{1}{\|w\|} \min_{n=1,\dots,N} \left(t_n (w^T x_n + w_0)\right)$

$= \underset{w, w_0}{\arg\max} \frac{1}{\|w\|} = \underset{w, w_0}{\arg\min} \frac{1}{2}\|w\|^2$ assuming for the closest point x_K $t_K(w^T x_K + w_0) = 1$

Solution: $w^* = \sum_{n=1}^{N} a_n^* t_n x_n$, a_n^* -> Lagrange multipliers

Support Vector: $SV = \{x_K \in D \mid t_K y(x_K) = 1\}$

Solution depends only on SV -> $y(x) = \sum_{x_j \in SV} a_j^* t_j x^T x_j + w_0^* = 0$

$w_0^* = \frac{1}{|SV|} \sum_{x_K \in SV} \left(t_K - \sum_{x_j \in S} a_j^* t_j x_K^T x_j\right)$

- SVM with ξ

Problem: $w^*, w_0^* = \underset{w, w_0}{\arg\min} \left[\frac{1}{2}\|w\|^2 + C \sum_{n=1}^{N} \xi_n\right]$

Solution: $w^* = \sum_{n=1}^{N} a_n^* t_n x_n$

Regression

Model : $y(x;w) = w_0 + w_1 x_1 + \dots + w_m x_m = w^T x$

$y(x;w) = w^T \phi(x)$ non linear in x, linear in w

Error function = $E_D(w) = \frac{1}{2} \sum_{n=1}^{N} [t_n - w^T \phi(x_n)]^2$ to minimize

Update : $w^{(\tau+1)} = w^{(\tau)} - \eta \nabla E_n$ after presentation of pattern n

$= w^{(\tau)} + \eta (t_n - w^{(\tau)T} \phi_n) \phi_n$

Regularization : $\min E_D(w) + \lambda E_W(w)$, $E_W(w) = \frac{1}{2} w^T w$

Kernel

Linear model with kernel : $y(x, \hat{w}) = \sum_{i=1}^{N} \alpha_i K(x, x')$ considering regularization

Solution is : $\alpha = (K + \lambda I_N)^{-1} t$, $K = X^T X$

Cost functions

Regression → linear output unit → mean squared error → $E(\theta) = \frac{1}{2} \sum_n^N (t_n - \theta^T x_n)^2$

Binary class → sigmoid " " → binary cross-entropy → $E(\theta) = -\ln P(t|x)$

$= \text{softplus}((1-2t)a)$

Multi class → softmax " " → categorical "

→ $E(\theta)_j = \ln \sum_j e^{a_j} - a_j$

NN layers dimensions and parameters

$$W_L = \frac{W_i - W_k + 2\text{padding}}{\text{stride}} + 1 \quad ; \quad h_L = \frac{h_i - h_k + 2\text{padding}}{\text{stride}} + 1$$

params = $(m \cdot m \cdot (l) + 1)(K)$ input output

× FC : # params = $(l+1)K$

Probabilistic Discriminative models

Objective : $P(C_k | \tilde{x}, D)$

with max likelihood : $\tilde{w}^* = \arg\max_{\tilde{w}} \ln P(t | \tilde{w}, x)$

For a logistic regression problem

Likelihood : $p(t|\tilde{w}) = \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{(1-t_n)}$

with : $y_n = p(C_1 | \tilde{x}_n) = \sigma(w^T \tilde{x}_n)$

Error function $\equiv -\ln p(t|\tilde{w}) = -\sum_{n=1}^{N} [t_n \ln y_n + (1-t_n) \ln(1-y_n)] = E(\tilde{w})$

Solution concept : $\tilde{w}^* = \arg\min_{\tilde{w}} E(\tilde{w})$

Solve with iterative Newton-Raphson method (gradient)

Gaussian Mixture Model

Gaussian $\mathcal{N}(x; \mu_k, \Sigma_k)$

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k)$$

With latent variables

z is a k dimensional binary random variable in which only one particular element z_k is equal to 1 and all the others are equal to 0

Marginal distribution over z: $p(z_k = 1) = \pi_k$ → Mixing coefficient

Thus: $p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$

Conditional distribution of x given a z:

$$p(x|z_k=1) = \mathcal{N}(x; \mu_k, \Sigma_k)$$

$$p(x|z) = \prod_{k=1}^{K} \mathcal{N}(x; \mu_k, \Sigma_k)^{z_k}$$

$p(x, z) = p(z)\, p(x|z)$

Marginal distribution of x: $p(x) = \sum_z p(z) p(x|z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k)$

Conditional probability of z given x:

$$\gamma(z_k) = p(z_k = 1 | x) = \frac{p(z_k=1)\, p(x|z_k=1)}{p(x)} = \frac{\pi_k \mathcal{N}(x; \mu_k, \Sigma_k)}{\sum_{j=1}^{J} \pi_j \mathcal{N}(x; \mu_j, \Sigma_j)}$$

π_k prior probability of $z_k = 1$

Expectation Maximization (Edoardo Montecchiani)

It uses maximum likelihood: $\underset{\mu, \Sigma, \pi}{\operatorname{argmax}} \ln p(x | \mu, \Sigma, \pi)$

E step: use current values for the parameters to evaluate the posterior probability (γ)

M step: use these probabilities to re-estimate the μ, Σ, π

Can be generalized to other distributions (not only Gaussians)

At maximum

$$\left.\begin{aligned} \mu_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n \\ \Sigma_k &= \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) (x_n - \mu_k)(x_n - \mu_k)^T \\ \pi_k &= \frac{N_k}{N} \end{aligned}\right\} \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

PCA

Problem: $\max_{u_i} u_i^T S u_i$ $\quad S$ = covariance matrix $= \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T = \frac{1}{N} X^T X$

$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$ $\quad = \begin{bmatrix} (x_1 - \bar{x})^T \\ \vdots \\ (x_N - \bar{x})^T \end{bmatrix}$

Problem constraint: $u_i^T u_i = 1$

Solution: $u_i^T S u_i = \lambda_i \qquad \forall i \in M$ reduced dimension

Projected point $\Rightarrow \langle u_1^T x_n, u_2^T x_n \dots u_M^T x_n \rangle$

Reconstructed point $\Rightarrow (u^T x) u^T$?

RL

$MDP = \langle X, A, \delta, r \rangle \quad \delta : X \times A \to X \; ; \; r \; X \times A \to \mathbb{R}$

Markov property: $x_{t+1} = \delta(x_t, a_t)$, $r_t = r(x_t, a_t)$

Policy: $\pi : X \to A$

Value function: $V^\pi(x_1) = r_1 + \gamma r_2 + \gamma^2 r_3 + \dots$

Non deterministic value fun $= E[V^\pi(x_1)]$

Optimal policy: $\pi^* = \arg\max_{\pi} V^\pi(x) \quad \forall x \in X$

Value iteration alg.: $\pi^* = \arg\max_{a \in A} [r(x,a) + \gamma V^*(\delta(x,a))]$

Deterministic Q-fun: $Q(x_t, a_t) = r(x_t, a_t) + \gamma \max_{a' \in A} Q(x_{t+1}, a')$

Non-deterministic " : $\hat{Q}_n(x,a) = \hat{Q}_{n-1}(x,a) + \alpha\left[r + \gamma \max_{a' \in A} \hat{Q}(x', a') - \hat{Q}_{n-1}(x,a)\right]$

$HMM = \langle X, Z, \pi_0 \rangle \quad Z \to$ observation model $P(z_t \mid x_t)$; π_0 initial distrib. $P(x_0)$

$POMDP = \langle X, A, Z, \delta, r, o \rangle$

$\delta(x', a, x') = P(x' \mid x', a)$ distribution over transition

Z is the set of observations

$o(x', a, z') = P(z' \mid x', a)$ prob. distribution over observations