Statistical Machine Learning – Exercise 2 – Michael Meidlinger

1 Bayes Classifier

a) We have $p(x,y) = p(y|x)p(x) = p(x|y)p(y)$ (1)

- c_1 is NOT equivalent. In fact, $p(x)$ is independent of y so that $\arg\max_y p(x)$ can be defined to output any arbitrary number
- c_2 is NOT equivalent. The output is independent of x always the least likely label
- c_3 is equivalent because of (1) and the fact that $p(x)$ is positive and independent of y
- c_4 is NOT equivalent, since $p(x|y) = p(y|x)\frac{p(x)}{p(y)}$

b) Using a similar argumentation as above we have

$c_7, c_8, c_9, c_{10}, c_{12}$ equivalent, the rest not

2 Gaussian Discriminant Analysis

a)
$$c(x) = \arg\max_{y\in\mathcal{Y}} p(y|x) = \arg\max_{y\in\mathcal{Y}} p(x|y)\,p(y) \overset{y\in\{-1,1\}}{=} \operatorname{sign}\left(\log\frac{p(x|1)p(y=1)}{p(x|-1)p(y=-1)}\right)$$

$$= \operatorname{sign}\left(\log\left(\exp\left(-\tfrac{1}{2}(\underline{x}-\underline{\mu}_1)^T\underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu}_1) + \tfrac{1}{2}(\underline{x}-\underline{\mu}_{-1})^T\underline{\underline{\Sigma}}^{-1}(\underline{x}-\underline{\mu}_{-1}) + \log\frac{p(y=1)}{p(y=-1)}\right)\right)\right)$$

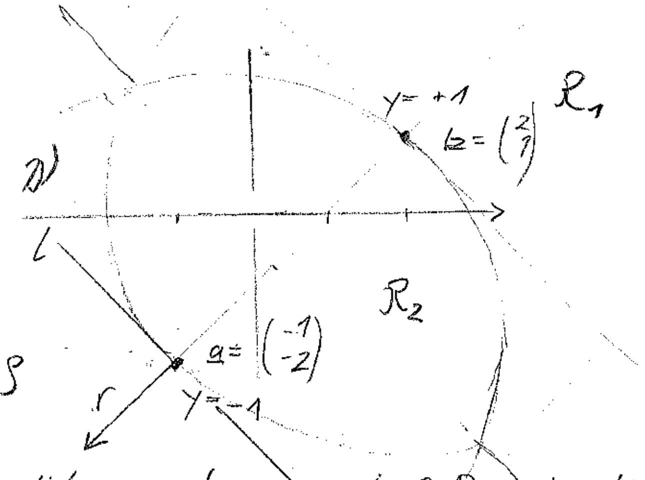
quadratic $\underline{x}$ terms cancel, quadratic $\underline{\mu}$ terms can be dropped (independent of y)

$$= \operatorname{sign}\Big(\underbrace{\underline{x}^T\underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 - \underline{x}^T\underline{\underline{\Sigma}}^{-1}\underline{\mu}_{-1}}_{\underline{x}^T(\underbrace{\Sigma^{-1}\mu_1 - \Sigma^{-1}\mu_{-1}}_{:=\underline{w}})} + \underbrace{\log\frac{p(y=1)}{p(y=-1)}}_{\theta}\Big)$$

$$= \operatorname{sign}(\langle\underline{x},\underline{w}\rangle + \theta) \qquad \text{q.e.d.}$$

b) With a few examples, the exact distributions cannot be estimated precisely. Thus one has to resort to parametric models of the distribution.

3 Robustness of the Perceptron



The robustnes ρ of a classifier g (with respect to $\mathcal{D}$) is the largest perturbation by which we can perturb the training samples without changing the predictions of g

$$g(x^i + \varepsilon) = g(x^i) \quad \forall i \quad \forall \varepsilon : \|\varepsilon\| < \rho$$

For a linear classifier, ρ is the smallest distance of any $x^i \in \mathcal{D}_x$ to the decision boundary. For the specified training data with points $x^i \in \mathbb{R}^2$ we can distinguish the following cases:

a.) $\begin{pmatrix} a \\ b \end{pmatrix}$ is an element of the ray $r \Rightarrow \mathcal{D}_{\binom{a}{b}}$ is not linearly separable and the perceptron algorithm won't converge

b.) $\begin{pmatrix} a \\ b \end{pmatrix}$ is an element of $R_2 := \left\{ \underline{x} \in \mathbb{R}^2 \,\middle|\, \|\underline{x} - \underline{a}\| < \frac{\|\underline{b} - \underline{a}\|}{2} \wedge \|\underline{x} - \underline{b}\| < \frac{\|\underline{b} - \underline{a}\|}{2} \right\}$

For that case, ρ is given by

$$\rho = \frac{\|\underline{x} - \underline{a}\|}{2} = \frac{1}{2} \left\| \begin{pmatrix} a + 1 \\ b + 2 \end{pmatrix} \right\| = \frac{1}{2} \sqrt{(a+1)^2 + (b+2)^2}$$

c.) $\binom{a}{b}$ is in $R_1 \Rightarrow \rho = \frac{1}{2} \frac{\|\underline{b} - \underline{a}\|}{2} = \frac{1}{2} \sqrt{\underbrace{3^2 + 3^2}_{2 \cdot 9}} = \frac{3}{\sqrt{2}}$

d.) $\binom{a}{b} \in \mathbb{R}_2 \setminus \{R_1 \cup R_2 \cup r\}$

The decision boundary is the line $d: \frac{\underline{x}_0 + \underline{x}}{2} + \lambda (\underline{b} - \underline{x})$ where

$$\underline{x}_0 = \underset{\underline{v} \in \{\underline{a}, \underline{b}\}}{\arg\min} \|\underline{v} - \underline{x}\| \quad \Rightarrow \quad \rho = \frac{\left(\frac{\underline{x} - \underline{x}_0}{2}\right) \times (\underline{b} - \underline{x})}{\|\underline{b} - \underline{x}\|} = \Bigg\{$$

- ρ_{max} is thus given by $3/\sqrt{2}$

4 Perceptron Training as Convex Optimization

Comparing Algorithm 1 with "Stochastic Gradient Descent" (ml 2014_09.pdf) we identify:

$$w_{t+1} \leftarrow w_t - \eta_t \underbrace{v}_{-n\nabla f_i(w_t) \text{ with random } i} \iff w_{t+1} \leftarrow w_t + y x$$

$$\hookrightarrow \quad -\eta_t n \nabla_w f_i(w_t) = \begin{cases} y^i x^i, & y\langle w_t, x\rangle \leq 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\Rightarrow \eta_t = \text{const} = 1 \qquad , \quad [x]^+ = \begin{cases} x, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$f_i(w) = \frac{1}{n}\left[\cancel{1/0} - y^i\langle w, x^i\rangle\right]^+ \Rightarrow f(w) = \sum_{i=1}^{n} f_i(w) = \overbrace{\frac{1}{n}\sum_{i=1}^{n}\left[\cancel{1/0} - y^i\langle w, x^i\rangle\right]^+}^{\text{Average of inner products is the cost function to be minimized}}$$

Advantages:

Shortcomings:

5 Hard-Margin SVM Dual

$$\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 \quad \text{subject to} \quad y^i(\langle w, x^i\rangle + b) \geq 1$$

We compute the Lagrangian:

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}\|w\|^2 + \underbrace{\sum_i \alpha_i\left(1 - y^i(\langle w, x^i\rangle + b)\right)}_{\sum_i \alpha_i - \sum_i \alpha_i y^i\langle w, x^i\rangle - b\sum_i \alpha_i y^i}, \quad h(\alpha) = \min_{(w,b)} \mathcal{L}(w,b,\alpha)$$

$$\Rightarrow \quad 0 = \partial/\partial w\, \mathcal{L} = w - \sum_i \alpha_i y_i x^i \Rightarrow w = \sum_i \alpha y^i x^i$$

$$0 = \partial/\partial b\, \mathcal{L} = \sum_i \alpha_i y^i$$

$\hookrightarrow$ Insert back:

$$h(\alpha) = \frac{1}{2}\|\sum_i \alpha_i y^i x^i\|^2 + \sum_i \alpha_i - \overbrace{\sum_i \alpha_i y^i \langle \sum_j \alpha_j y^j x^j, x^i\rangle}^{\|\sum_i \alpha_i y^i x^i\|^2}$$

$$= -\frac{1}{2}\|\sum_i \alpha_i y^i x^i\|^2 + \sum_i \alpha_i$$

$$= -\frac{1}{2}\langle \sum_i \alpha_i y^i x^i, \sum_j \alpha_j y^j x^j\rangle + \sum_i \alpha_i = -\frac{1}{2}\sum_i\sum_j \alpha_i \alpha_j y^i y^j \langle x^i, x^j\rangle + \sum_i \alpha_i$$

Dual Problem: $\min_{\alpha > 0} h(\alpha)$ subject to

$$\sum_i \alpha_i y^i = 0 \quad ,$$

6 Missing Proofs

a.) We need to show $\underset{k}{f}(w) \geq \underset{k}{f}(w_0) + \langle v, w-w_0 \rangle \Longrightarrow f(w) \geq f_k(w_0) + \langle v, w-w_0 \rangle \;\; \forall k$, where $f(w) = \max\{f_1(w), \dots, f_K(w)\}$

↘ Use assumption

$$f(w) = \max_{k'=1,\dots,K} f_{k'}(w) \geq f_K(w) \geq \underbrace{f_K(w_0)}_{=f(w_0) \text{ (assumption)}} + \langle v, w-w_0 \rangle = f(w_0) + \langle v, w-w_0 \rangle \qquad \square$$

b.)
$$\| \overset{w_t - \eta v}{w_{t+1}} - w^* \|^2 = \|w_{t+1}\|^2 + \|w^*\|^2 - 2 \langle w_{t+1}, w^* \rangle - \|w_t - w^*\|^2$$

$$= \|w_t\|^2 + \|w^*\|^2 - 2 \langle w_t, w^* \rangle + 2\eta \langle v, w^* \rangle + \eta^2 \|v\|^2 \qquad (1)$$

$$\|w_t - w^*\|^2 = \|w_t\|^2 + \|w^*\|^2 - 2 \langle w_t, w^* \rangle \leq \|v\| \, \|w_t - w^*\| \qquad (2)$$

$$v \in \partial f(w_t) \Longleftrightarrow f(w) \geq f(w_t) + \langle v, w - w_t \rangle \overset{w=w^*}{\Longleftrightarrow} \langle v, w_t - w^* \rangle \geq f(w_t) - f(w^*) \qquad (3)$$

f is convex $\Longrightarrow$ $f(\theta w_1 + (1-\theta) w_2) \leq \theta f(w_1) + (1-\theta) f(w_2)$

$$\overset{w=w_{t+1}}{\Longleftrightarrow} f(w_t) - f(w_{t+1}) \leq \eta \|v\|^2$$

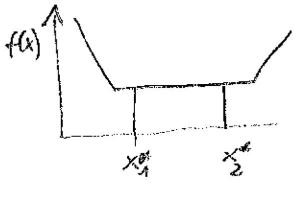
$$\hookrightarrow \|w_{t+1} - w^*\|^2 = \|w_t - \eta v - w^*\| = \|w_t - w^*\|^2 - 2\eta \overbrace{v^T (w_t - w^*)}^{\langle v, w_t - w^* \rangle} + \eta^2 \|v\|^2$$

$$\overset{(3)}{\leq} \|w_t - w^*\|^2 \underbrace{- 2\eta \left(f(w_t) - f(w^*) \right) + \eta^2 \|v\|^2}$$

$$< 0 \text{ iff } \eta^2 \|v\|^2 < 2\eta \left(\underbrace{f(w_t) - f(w^*)}_{\geq 0} \right) \Longleftrightarrow \eta < \frac{2(f(w_t) - f(w^*))}{\|v\|^2}$$

$$< \|w_t - w^*\| \quad \text{if} \quad \eta < \frac{\overbrace{2(f(w_t) - f(w^*))}^{\geq 0}}{\|v\|^2} \qquad \square$$

c.) A convex function can have multiple minima only if it is not strictly convex (i.e. $f(\theta w_1 + (1-\theta) w_2) < \theta f(w_1) + (1-\theta) f(w_2)$ is not guaranteed $\forall w_1, w_2, \theta \in [0,1]$). If this is the case, multiple global minima can occur, but they are "next to each other", i.e. if x_1^* is a minimum point and $f(x)$ has multiple minima then any other minima x_2^* will be in a ball around x_1^*: $x_1^* = \arg\min_x f(x) \wedge x_2^* = \arg\min_x f(x)$



$\Longrightarrow$ any convex combination $\theta x_1^* + (1-\theta) x_2^*$ will also be a minimum

d.) Since $g(\alpha)$ is the pointwise maximum of a set of convex functions (in fact $g(\alpha) = \max_\theta h_\theta(\alpha)$ where $h_\theta(\alpha) = f(\theta) + \sum_{i=1}^{k} \alpha_i g_i(\theta)$ is <u>affine</u> in α), $g(\alpha)$ is convex in α. For a proof, see Boyd/Vandenberghe Sec. 3.2.3