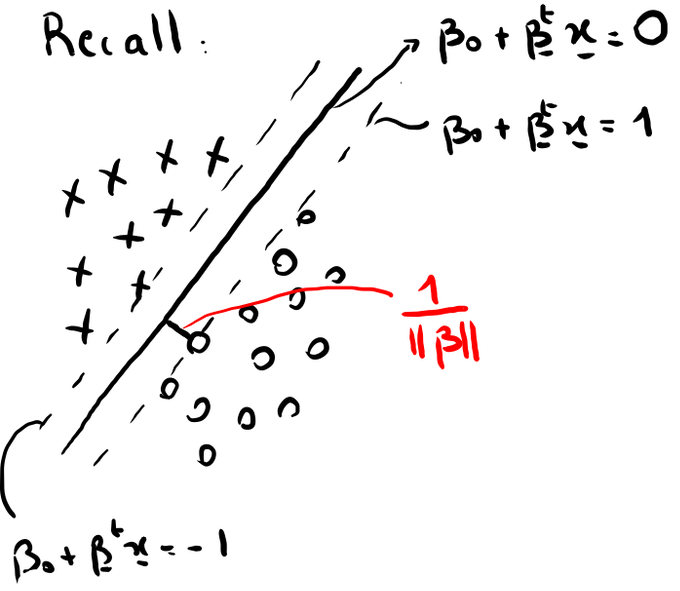
26/4/21

Recall:



SUPPORT VECTOR CLASSIFIER

$$\min_{\beta_0, \beta, \xi} \frac{1}{2}\|\beta\|^2 \rightarrow \beta_1^2 + \beta_2^2 + \dots + \beta_p^2$$

QUADRATIC CONSTRAINED OPTIMIZATION

s.t.

C_1 $y_i(\beta_0 + \beta^T x_i) \geq 1 - \varepsilon_i$, $i = 1, \dots, n$

C_2 $\varepsilon_i \geq 0$ $i = 1, \dots, n$

C_3 $\sum_{i=1}^{n} \varepsilon_i \leq C$ → tuning parameter

A BRIEF INTRODUCTION TO OPTIMIZATION

Some examples: Multinomial, Lasso, Ridge

Multinomial ↑ equality constraints $\sum_{j=1}^{k} \theta_j = 1$

Lasso ↑ inequality constraints $\sum |\beta_j| \leq C$

Ridge ↑ $\sum \beta_j^2 \leq C$

Generic optimisation problem under equality constraints:

$$\left.\begin{aligned} \min_{\underline{\alpha}} \quad & f(\underline{\alpha}) \\ \text{s.t.} \quad & h_i(\underline{\alpha}) = 0 \;, \quad i = 1, \dots, n \end{aligned}\right)$$

EXAMPLE Multinomial

Log-lik. $\ell(\underline{\theta}) = \log(n!) - \sum_{i=1}^{G} \log(x_i!) + \sum_{i=1}^{G} x_i \log(\theta_i)$

parameters: $\theta_1, \dots, \theta_G$

$\max \ \ell(\underline{\theta})$ ← $f(\underline{\alpha})$

s.t. $\sum_{i=1}^{G} \theta_i = 1$

$\Leftrightarrow \underbrace{1 - \sum_{i=1}^{G} \theta_i}_{} = 0$ ← $h_2(\underline{\alpha})$

LAGRANGIAN

$$\ell(\underline{\alpha}, \underline{\gamma}) = f(\underline{\alpha}) + \underbrace{\sum_{i=1} \gamma_i h_i(\underline{\alpha})}_{\gamma_1 h_1(\underline{\alpha}) + \gamma_2 h_2(\underline{\alpha}) + \dots}$$

UNCONSTRAINED OPTIMIZATION

with $\underline{\gamma} = (\gamma_1, \dots, \gamma_n)$ Lagrange multipliers

SOLUTION $\dfrac{\partial \ell(\underline{\alpha}, \underline{\gamma})}{\partial \underline{\alpha}} = 0 \;, \quad \dfrac{\partial \ell(\underline{\alpha}, \underline{\gamma})}{\partial \underline{\gamma}} = 0$ } $\hat{\underline{\alpha}}, \hat{\underline{\gamma}}$

$\Rightarrow \hat{\underline{\alpha}}$ is the solution of the original constrained optimisation problem

$$\underbrace{\frac{\partial \ell(\underline{\alpha}, \underline{\gamma})}{\partial \gamma_1} = h_1(\underline{\alpha}) = 0}$$

↑ $\underline{\alpha}$ satisfies constraint 1

EXAMPLE: Multinomial

Lagrangian

$$\ell^*(\underline{\theta}, \lambda) = \log(n!) - \sum_{i=1}^{G} \log(x_i!) + \sum_{i=1}^{G} x_i \log(\theta_i) + \lambda \left(1 - \sum_{i=1}^{G} \theta_i\right)$$

$$\frac{\partial \ell^*}{\partial \underline{\theta}} = 0 \qquad \frac{\partial \ell^*}{\partial \lambda} = 0$$

$$\frac{\partial \ell^*}{\partial \theta_i} = \frac{x_i}{\theta_i} - \lambda = 0 \quad , i = 1, \dots, G \qquad \frac{\partial \ell^*}{\partial \lambda} = \left(1 - \sum_{i=1}^{G} \theta_i\right) = 0$$

$$\theta_i = \frac{x_i}{\lambda} \qquad \sum_{i=1}^{G} \theta_i = 1$$

$$\sum_{i=1}^{G} \frac{x_i}{\lambda} = 1 \quad \Rightarrow \quad \hat{\lambda} = \sum_{i=1}^{G} x_i = n$$

$$\Rightarrow \quad \boxed{\hat{\theta}_i = \frac{x_i}{n}}$$

OBSERVED FREQUENCIES

Let's generalise to the case of INEQUALITY CONSTRAINTS

(PRIMAL) PROBLEM

$$\min_{\underline{\alpha}} \; f(\underline{\alpha})$$

s.t.

$$h_i(\underline{\alpha}) = 0 \quad i = 1, \dots, p \quad \text{(equality constraints)}$$

$$g_i(\underline{\alpha}) \le 0 \quad i = 1, \dots, n \quad \text{(inequality constraints)}$$

GENERALIZED LAGRANGIAN

$$\ell(\underline{\alpha}, \underline{\gamma}, \underline{\delta}) = f(\underline{\alpha}) + \sum_{i=1}^{p} \gamma_i h_i(\underline{\alpha}) + \sum_{i=1}^{n} \delta_i g_i(\underline{\alpha})$$

with $\gamma_1, \dots, \gamma_p, \delta_1, \dots, \delta_n$ Lagrange multipliers

To justify the use of this function consider

$$\theta_p(\underline{\alpha}) = \max_{\underline{\gamma}, \underline{\delta}, \delta_i \ge 0} \ell(\underline{\alpha}, \underline{\gamma}, \underline{\delta})$$

$$\min_{\underline{\alpha} \text{ constraints}} f(\underline{\alpha}) = \min_{\underline{\alpha}} \theta_p(\underline{\alpha}) = \min_{\alpha} \left(\max_{\underline{\gamma}, \underline{\delta}, \delta_i \ge 0} \ell(\underline{\alpha}, \underline{\gamma}, \underline{\delta}) \right)$$

Take $\underline{\alpha}$ that does not satisfy the constraints ($h_i(\underline{\alpha}) \neq 0$ for some i, or $g_i(\underline{\alpha}) > 0$ for some i)

then $\max_{\underline{\gamma}, \underline{\delta}, \delta_i \ge 0} \ell(\underline{\alpha}, \underline{\gamma}, \underline{\delta}) = +\infty$

So the minimum must be achieved by an α that satisfies the constraints.

Take $\underline{\alpha}$ that satisfies the constraints. Then $\sum \gamma_i h_i(\underline{\alpha}) = 0$, $g_i(\underline{\alpha}) \le 0$

Then take $\delta_i = 0 \Rightarrow \ell(\underline{\alpha}, \gamma, \underline{\delta}) = f(\underline{\alpha})$

PRIMAL PROBLEM

$$\min_{\substack{\underline{\alpha} \\ \text{constraints}}} f(\underline{\alpha}) = \min_{\underline{\alpha}} \theta_P(\underline{\alpha}) = \min_{\underline{\alpha}} \max_{\substack{\underline{\gamma}, \underline{\delta} \\ \delta_i \geq 0}} \ell(\underline{\alpha}, \underline{\gamma}, \underline{\delta})$$

DUAL PROBLEM

$$\max_{\substack{\underline{\gamma}, \underline{\delta} \\ \delta_i \geq 0}} \theta_D(\underline{\gamma}, \underline{\delta}) = \max_{\substack{\underline{\gamma}, \underline{\delta} \\ \delta_i \geq 0}} \min_{\underline{\alpha}} \ell(\underline{\alpha}, \underline{\gamma}, \underline{\delta})$$

These two problems in general do not have the same solution.

One can show:

$$\max - \min \leq \min - \max$$

But they are the same in the case of f and g_i convex functions and h_i linear, then

$$\max - \min = \min - \max$$

KARUSH-KUHN-TUCKER (KKT) CONDITIONS

We assume $\theta_p(\underline{\alpha}^*) = \theta_D(\underline{\gamma}^*, \underline{\delta}^*)$. Then

$$\delta_i^* g_i(\underline{\alpha}^*) = 0 \qquad i = 1, \dots, n$$

COMPLEMENTARY SLACKNESS CONDITIONS

Proof: Let $\underline{\alpha}^*, \underline{\gamma}^*, \underline{\delta}^*$ be the optimum.

$$f(\underline{\alpha}^*) = \min_{\underline{\alpha}} l(\underline{\alpha}, \underline{\gamma}^*, \underline{\delta}^*) = \min_{\underline{\alpha}} \left(f(\underline{\alpha}) + \sum_{i=1}^{p} \gamma_i^* h_i(\underline{\alpha}) + \sum_{i=1}^{n} \delta_i^* g_i(\underline{\alpha}) \right)$$

$$\leq f(\underline{\alpha}^*) + \underbrace{\sum_{i=1}^{p} \gamma_i^* h_i(\underline{\alpha}^*)}_{=0} + \underbrace{\sum_{i=1}^{n} \delta_i^* g_i(\underline{\alpha}^*)}_{\leq 0}$$

$$\leq f(\underline{\alpha}^*)$$

$$\Rightarrow \sum_{i=1}^{n} \delta_i^* g_i(\underline{\alpha}^*) = 0 \iff \delta_i^* g_i(\underline{\alpha}^*) = 0 \qquad i = 1, \dots, n$$

$$\delta_i^* g_i(\underline{\alpha}^*) \leq 0$$

$$\left.\begin{array}{l} \delta_i^* > 0 \Rightarrow g_i(\underline{\alpha}^*) = 0 \\ g_i(\underline{\alpha}^*) < 0 \Rightarrow \delta_i^* = 0 \end{array}\right\}$$

Back to **SUPPORT VECTOR CLASSIFIER**

$$\min_{\beta_0,\underline{\beta},\underline{\xi}} \ \frac{1}{2}\|\underline{\beta}\|^2 \qquad \beta_1^2+\beta_2^2+\dots+\beta_p^2 \quad \text{(convex)}$$

s.t.

$$y_i(\beta_0+\underline{\beta}^T\underline{x}_i) \geq 1-\xi_i, \quad i=1,\dots,n$$
$$\xi_i \geq 0$$
$$\sum_{i=1}^n \xi_i \leq C$$

constraints: convexity

$$C_1: -[y_i(\beta_0+\underline{\beta}^T\underline{x}_i)-1+\xi_i] \leq 0$$
$$C_2: -\xi_i \leq 0$$
$$C_3: \sum_{i=1}^n \xi_i - C \leq 0$$

LAGRANGIAN

$$l(\beta_0,\underline{\beta},\underline{\xi},\underline{\gamma},\underline{\delta},\mu) = \frac{1}{2}\|\underline{\beta}\|^2 - \sum_{i=1}^n \gamma_i[y_i(\beta_0+\underline{\beta}^T\underline{x}_i)-1+\xi_i] - \sum_{i=1}^n \delta_i\xi_i + \mu\left(\sum_{i=1}^n \xi_i - C\right)$$

($\|\underline{\beta}\|^2 = \beta_1^2+\beta_2^2+\dots+\beta_p^2$; $\sum \gamma_i y_i \beta^T x_i$)

for $\gamma_i \geq 0, \delta_i \geq 0, \mu \geq 0$, $i=1,\dots,n$

Then

$$\min_{\beta_0,\underline{\beta},\underline{\xi}\ \text{constraints}} \frac{1}{2}\|\underline{\beta}\|^2 = \underbrace{\min_{\beta_0,\underline{\beta},\underline{\xi}} \max_{\underline{\gamma},\underline{\delta},\mu \geq 0} l(\cdot)}_{\text{PRIMAL PROBLEM}} \overset{\text{convex}}{\geq} \underbrace{\max_{\underline{\gamma},\underline{\delta},\mu \geq 0} \min_{\beta_0,\underline{\beta},\underline{\xi}} l(\cdot)}_{\text{DUAL PROBLEM}}$$

Taking derivatives:

$$\frac{\partial l}{\partial \beta_0} = -\sum_{i=1}^n \gamma_i y_i = 0 \Rightarrow \sum_{i=1}^n \gamma_i y_i = 0$$

$$\frac{\partial l}{\partial \underline{\beta}} = \underline{\beta} - \sum_{i=1}^n \gamma_i y_i \underline{x}_i = 0 \Rightarrow \boxed{\underline{\beta} = \sum_{i=1}^n \gamma_i y_i \underline{x}_i}$$

$$\frac{\partial l}{\partial \xi_i} = -\gamma_i - \delta_i + \mu = 0 \Rightarrow \gamma_i = -\delta_i + \mu, \quad i=1,\dots,n$$

$0 \leq \delta_i = \mu - \gamma_i$

Resubstituting back in the objective function gives:

$$\Theta_D(\underline{\gamma}, \underline{\delta}, \mu) = \frac{1}{2}\underline{\beta}^t\underline{\beta} - \sum_{i=1}^n (\gamma_i y_i \beta_0 + \gamma_i y_i \underline{\beta}^t \underline{x}_i - \gamma_i + \gamma_i \varepsilon_i) - \sum_{i=1}^n \delta_i \varepsilon_i + \mu\left(\sum_{i=1}^n \varepsilon_i - C\right)$$

$$= \frac{1}{2}\underline{\beta}^t\underline{\beta} - \underline{\beta}^t\underline{\beta} + \sum_{i=1}^n \gamma_i - \mu C$$

$$= \sum_{i=1}^n \gamma_i - \frac{1}{2}\sum_{i,j=1}^n \gamma_i \gamma_j y_i y_j \underbrace{\underline{x}_i^t \underline{x}_j}_{\underline{x}_i \cdot \underline{x}_j \text{ scalar product}} - \mu C$$

$\underline{\beta} = \sum_{i=1}^n \gamma_i y_i \underline{x}_i$

$- \sum_{i=1}^n \gamma_i \varepsilon_i - \sum_{i=1}^n \delta_i \varepsilon_i + \mu \sum_{i=1}^n \varepsilon_i$

$= \sum_{i=1}^n (-\gamma_i - \delta_i + \mu)\varepsilon_i = 0$

DUAL PROBLEM

$$\max_{\underline{\gamma}, \mu} \quad \sum_{i=1}^n \gamma_i - \frac{1}{2}\sum_{i,j=1}^n \gamma_i \gamma_j y_i y_j \underline{x}_i \cdot \underline{x}_j \underbrace{- \mu C}$$

μC = constant

$\mu = \frac{\text{constant}}{C}$

$$\text{s.t.} \quad 0 \le \gamma_i \le \mu \qquad [\mu - \gamma_i \ge 0]$$

$$\sum_{i=1}^n \gamma_i y_i = 0$$

Solving the dual problem is equivalent to solving the original primal problem so

$$\underline{\beta} = \sum_{i=1}^n \gamma_i y_i \underline{x}_i$$

Note: $\underline{\gamma}$ has dimension n, so the solution is not too affected by a large p

Moreover, many of the γ_i will be zero.

Using the KKT conditions, at the optimum:

$$\delta_i^* g_i(\underline{\alpha}^*) = 0 \quad , i = 1, \dots, n$$

$$\gamma_i \left[y_i(\beta_0 + \underline{\beta}^T \underline{x}_i) - 1 + \varepsilon_i \right] = 0 \quad , \quad i = 1, \dots, n$$

Used to retrieve $\hat{\beta}_0$ by taking an average over $i = 1, \dots, n$

But more importantly, $\gamma_i \neq 0$ identifies the support vectors!

This is because if an observation is correctly classified (on the right side of margin)

then $\varepsilon_i = 0$ and $y_i(\beta_0 + \underline{\beta}^T \underline{x}_i) > 1 \Rightarrow y_i(\beta_0 + \underline{\beta}^T \underline{x}_i) - 1 + \varepsilon_i \neq 0 \Rightarrow \gamma_i = 0$

So $\gamma_i \neq 0$ identifies the points on the margin or misclassified.