Structural Risk Minimization

Empirical misclassification in Binary classification:

$$\mathcal{L}_n(\operatorname{sign}(h)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i \cdot \operatorname{sign}(h)(x_i) < 0}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i h(x_i) < 0}$$

Risk convexification: Replace $z \mapsto \mathbb{1}_{z<0}$

by some convex function $z \mapsto \ell(z)$

Remark (Logistic Regression) In some cases,

risk convexification can be interpreted as maximum

likelihood estimation of some Generalized Lin. Model

Consider a binary response $Y \in \{0,1\}$. In logistic regression, the log-likelihood ratio is given by

$$\log \frac{\mathbb{P}_2(Y=1 \mid X=x)}{\mathbb{P}_1(Y=0 \mid X=x)} = \beta_0 + \beta^T x$$

that yields

$$\underbrace{\mathbb{P}_1(Y=1 \mid X=x)}_{\mathbb{E}[Y \mid X=x]} = \frac{e^{\beta_0 + \beta^T x}}{1 + e^{\beta_0 + \beta^T x}}$$

The negative log-likelihood is:

$$-\frac{1}{n} \sum_{i=1}^{n} \left\{ y_i \log \mathbb{P}_1[Y=1 \mid x_i] + (1-y_i) \log \mathbb{P}_1[Y=0 \mid x_i] \right\}$$

In ML, it is more common to code

$$Y \in \{-1, +1\}$$

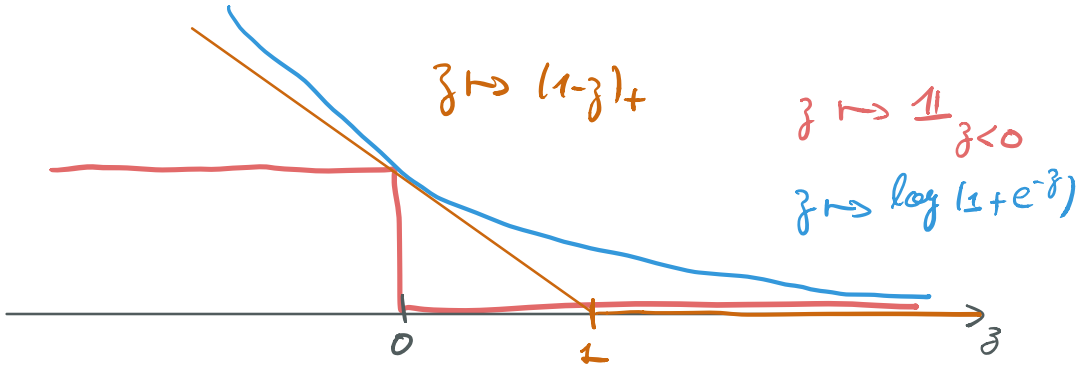
and, in this case, the negative log-likelihood is

$$\frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i h(x_i)}\right)$$

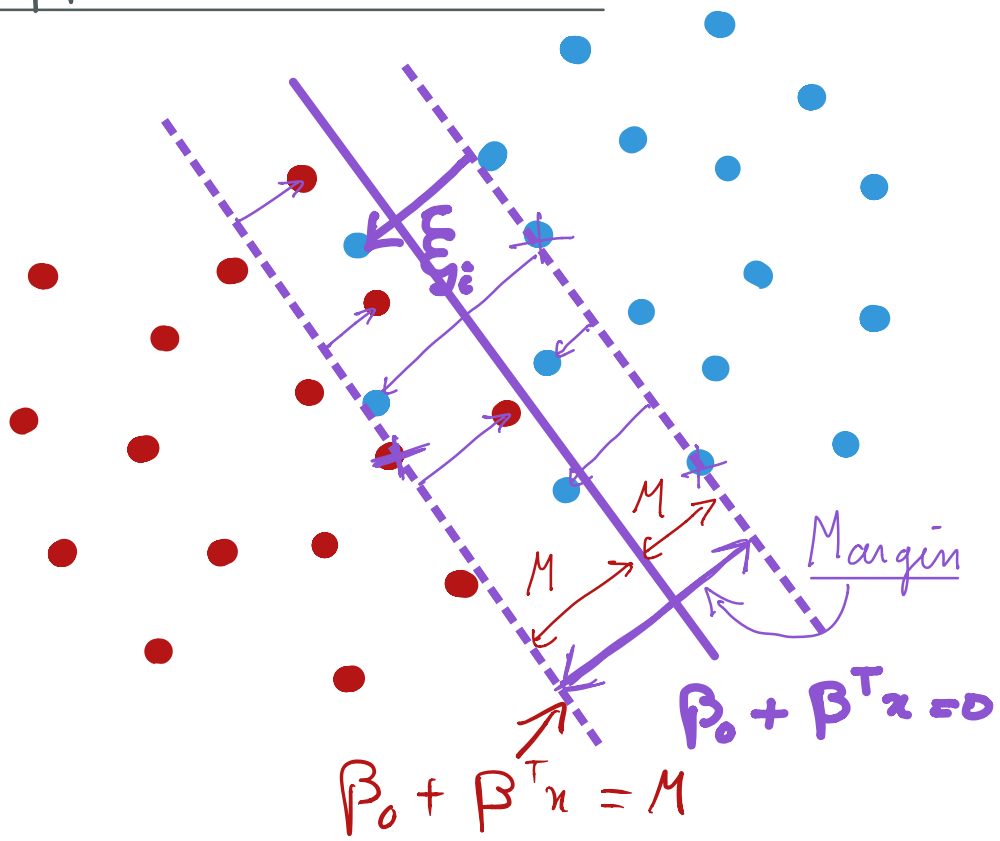
with $h(x_i) = \beta_0 + \beta^T x_i$

Popular loss functions

- logit loss : $\log(1 + e^{-z})$
- hinge loss : $(1-z)_+ := \max(0, 1-z)$



Support Vector Machines



$$l(z) = (1-z)_+ \qquad M = \frac{1}{\|\beta\|_2}$$

Maximize M

β_0, β, ξ_i

$$\underline{\text{subject to}}\ y_i(\beta_0 + \beta^T x_i) \geq M(1 - \xi_i)$$

$$\text{and} \quad \begin{cases} \xi_i \geq 0 \\ \sum_{i=1}^{n} \xi_i \leq C \\ \\ \|\beta\|_2 = 1 \end{cases}$$

This program is equivalent to:

$$\min_{\beta_0, \beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \underbrace{(1 - y_i h(x_i))_+}_{= \xi_i} + \lambda \|\beta\|_2^2 \right\}$$

where $h(x) = \beta_0 + \beta^T x$

solution is given by :

$$\hat{\beta} = \sum_{i=1}^{m} \hat{\alpha}_i \, y_i \, x_i$$

$$\hat{\beta}_0 = \sum_{i=1}^{m} \hat{\alpha}_i \, y_i$$

$$\hat{h}(x) = \sum_{i=1}^{m} \hat{\alpha}_i \, y_i \langle \begin{pmatrix} x_i \\ 1 \end{pmatrix}, \begin{pmatrix} x \\ 1 \end{pmatrix} \rangle$$

$$= \hat{\beta}_0 + \hat{\beta}^T x$$

$$X = \begin{pmatrix} x \\ 1 \end{pmatrix} \quad X_i = \begin{pmatrix} x_i \\ 1 \end{pmatrix} :$$

$$\hat{h}(x) = \sum_{i=1}^{n} \hat{\alpha}_i \, y_i \langle X_i, X \rangle$$

$k(x_i, x)$

Reproducing Kernel Hilbert Spaces

- In some ML problems, solutions are given by product of matrices:

$$\mathbb{X}_n^T \mathbb{X}_n \,, \quad \mathbb{X}_n \mathbb{Y}_n \ \dots$$

where $\mathbb{X}_n = [x_1 \cdots x_n]$

$\mathbb{Y}_n = [y_1 \cdots y_n]$

Idea: $\left(\mathbb{X}_n^T \mathbb{X}_n\right)_{ij} = \langle x_i, x_j \rangle$

$\left(\mathbb{X}_n \mathbb{Y}_n\right)_i = y_i x_i$

$$\langle x_i, x_j \rangle \longleftarrow k(x_i, x_j)$$

$$\langle y_i x_i, x \rangle \longleftarrow y_i\, k(x_i, x)$$

RKHS: $\forall n \geq 2, \ \forall x_1, x_2, \dots x_n \in \mathcal{X}$

$$K = \left[k(x_i, x_j) \right]_{1 \leq i,j \leq n} \succeq 0$$

- $k(x, y)$ symmetric :

$$k(x, y) = k(y, x)$$

ex : $k(x, y) = \langle x, y \rangle$

SVM → RKHS framework

- $\frac{1}{n} \sum_{i=1}^{n} (1 - y_i h(x_i))_+ + \lambda \|\beta\|_2^2$

- $\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \begin{pmatrix} x_i \\ 1 \end{pmatrix}$

$= \sum_{i=1}^{n} \hat{c}_i X_i$ $\qquad h(X) = X^T [Kc]$

- $\|\hat{\beta}\|_2^2 = \sum_{i,j} c_i c_j \langle x_i, x_j \rangle$

$= c^T K c$ $\qquad$ with $K = (\langle x_i, x_j \rangle)_{ij}$

- $h(X_i) = \langle \hat{\beta}, X_i \rangle = \sum_j c_j \langle X_j, X_i \rangle$

$$= \left[K c \right]_i$$

With $\begin{cases} K = \left[\langle x_i, x_j \rangle \right]_{i,j} \\ \beta = \sum c_i \begin{pmatrix} x_i \\ 1 \end{pmatrix} \end{cases}$

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} \left(1 - y_i \left[K c \right]_i \right)_+ + c^T K c \right\}$$

Kernel SVM

Instead of $K = \left[\langle x_i, x_j \rangle \right]_{i,j} = X_n^T X_n$

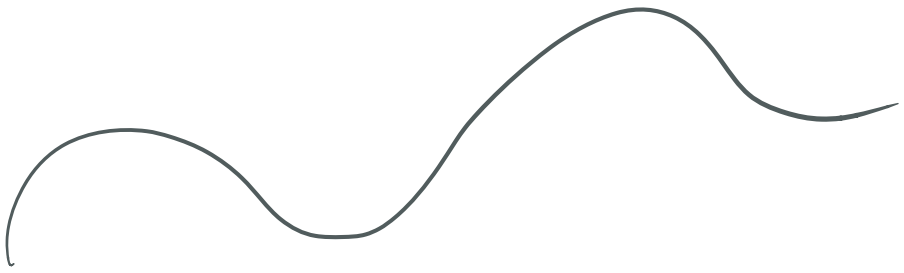
We can choose any RKHS matrix K.

$$K = \left[k(x_i, x_j) \right]_{i,j}$$

with k "good" kernel

$$h(x) = \sum_{i=1}^{m} c_i \, k(x_i, x)$$

$$\left\{ x : \; h(x) = 0 \right\} = \left\{ x : \; \sum_{i=1}^{m} c_i \, k(x_i, x) = 0 \right\}$$



We look at the kernalized SVM:

$$\hat{h}(\cdot) = \sum_{j=1}^{n} \hat{\beta}_j \, k(X_j, \cdot)$$

$$\hat{\beta} \in \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} (1 - Y_i \, h(X_i))_+ + \lambda \sum_{i,j} \beta_i \beta_j \, k(X_i, X_j) \right\}$$

Prop: $\hat{\beta}$ is such that:

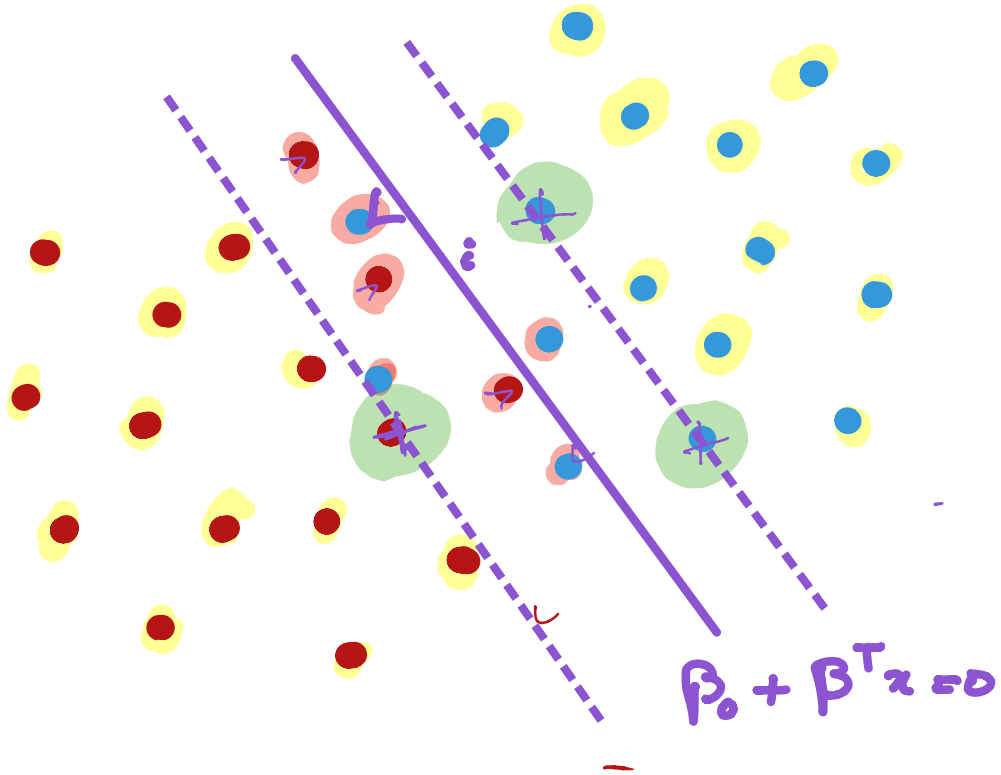
$$\hat{\beta}_i = 0 \quad \text{if} \quad Y_i \, \hat{h}(X_i) > 1$$

$$\hat{\beta}_i = \frac{Y_i}{2\lambda n} \quad \text{if} \quad Y_i \, \hat{h}(X_i) < 1$$

$\hat{\beta}_i$ is such that

$$0 \leq y_i \hat{\beta}_i \leq \frac{1}{2\lambda m}$$

if $y_i \hat{h}(x_i) = 1$



Proof: Slack variables:

$$\xi_i = \left(1 - Y_i \underbrace{[K\beta]_i}_{h(x_i)}\right)_+$$

Problem $\Leftrightarrow$

$$(\hat{\beta}, \hat{\xi}) \in \arg\min_{\beta, \xi} \left\{ \frac{1}{n} \sum_{i=1}^{n} \xi_i + \lambda \beta^T K \beta \right\}$$

$$\begin{cases} \xi_i \geq 1 - Y_i [K\beta]_i \\ \xi_i \geq 0 \end{cases}$$

KKT conditions for the Lagrangian:

$$\mathcal{L}(\beta, \xi, \alpha, \gamma) = \left\{ \frac{1}{n} \sum_i \xi_i + \lambda \beta^T K \beta \right.$$

$$- \sum_{i=1}^{n} \left(\alpha_i \left(\xi_i - 1 + y_i [K\beta]_i \right) + \gamma_i \xi_i \right) \Bigg\}$$

First order: $2\lambda [K\hat{\beta}]_j = \sum_i K_{ij} \alpha_i y_i$

and $\alpha_j + \gamma_j = \frac{1}{n}$

Slackness: $\min \left(\alpha_i, \hat{\xi}_i - 1 + y_i [K\hat{\beta}]_i \right) = 0$

$\min \left(\gamma_i, \hat{\xi}_i \right) = 0$