HW1 kernel SVM

Syllabus.pdf
- Seminar 20-minute
- final project (requires implementation)
  1-page proposal
  + presentation
  + report

Sign-up sheets:
seminar
+ final project presentations

Empirical risk minimization

loss function / risk : on data space $\mathcal{X}$

$$\ell((x,y), h) \in \mathbb{R}^+.$$

h ↓ hypothesis (function to be learned from data, to mimic / approx target)

$x_i, y_i = \{\pm 1\}$
x_i ↓ inputs, y_i ↑ labels

target ↓ $f(x) = y$

Want: find h s.t. $h(x) \approx y$.

Given: $(x_i, y_i = f(x_i))_{i \in [m]}$.

$(x_i, y_i) \sim \mathcal{D}$ (data distribution) probability measure

goal: Find an h
s.t. $\min_h \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell((x,y), h)$

Loss fn: $\ell((x,y), h) = 0$ if $f(x) = h(x) = y$.

Hypothesis class : function class where h lives.

Goal: $\min_{h \in \mathcal{H}} \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell((x,y), h)$
$\mathcal{H}$ ↓ hypothesis class

Examples of loss functions

(for binary classification)

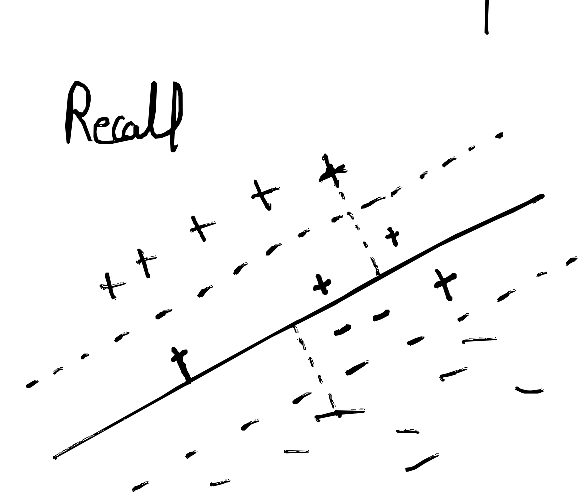
- $\ell((x,y), h) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{o.w.} \end{cases}$

zero-one loss.

$$\min_{h \in \mathcal{H}} \mathbb{E}_{(x,y) \sim \mathcal{D}} \mathbb{1}_{\{h(x) \neq y\}}$$

- 

Recall



$h(x) = (w^T x + b)$

Soft SVM:
Constraints:
$y_i(w^T x_i + b) \geq 1 - \xi_i$
$y_i h(x_i) \leq 1 - \xi_i$

obj fn: $\min \|\xi\|$

$$\ell((x,y), h) = \max\left\{0, 1 - \frac{y h(x)}{\rho}\right\}$$

Goal: $\min_{h \in \mathcal{H}} \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell((x,y), h)$

Hypothesis classes

Bin. Class: $h(x) = \text{sgn}(w^T x + b)$
SVM

parameterized by (w, b), $w \in \mathbb{R}^d$, $b \in \mathbb{R}$

Goal: $\min_{h \in \mathcal{H}} \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell((x,y), h)$: generalization error

Assumption: $(x_i, y_i) \sim \mathcal{D}$ iid

$\mathbb{E}_{(x,y) \sim \mathcal{D}} \ell((x,y), h)$: gene. err of h.

$\mathcal{D}$: unknown

Empirical risk minimization (ERM)

Actual : $\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell((x_i, y_i), h)$

$$\frac{1}{m} \sum_{i=1}^{m} \ell((x_i, y_i), h) \xrightarrow{m \to \infty} \mathbb{E}_{(x,y) \sim \mathcal{D}} \ell((x,y), h)$$
(LLN)

ML: Solving ERMs.

Supervised learning: data are of the form (x_i, y_i)
↑ y_i evaluations of the target function

unsupervised learning: data are of the form $\{x_i\}_{i=1}^m$

how to solve ERM
Optimization:

1. Parameterize $\mathcal{H}$
   e.g. NNs, kernels

2. w : set of parameters $\in \mathbb{R}^p$

ERM

$$\min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \ell((x_i, y_i), h) = \min_{w \in \mathbb{R}^p} \frac{1}{m} \sum_{i=1}^{m} \ell((x_i, y_i), h(x_i, w))$$

$x \to h(x, w)$: on input space
$w \to h(x, w)$: on parameter space

Overparameterize : $p \gg d \times m$

3. GD :

$$R_S(w) = \frac{1}{m} \sum_{i=1}^{m} \ell((x_i, y_i), h(x_i, w)) \quad \text{or } R_S(h)$$

R_S ↑ empirical risk

$S = \{(x_i, y_i)\}_{i \in [m]}$

GD Update} : $w_{t+1} = w_t - \eta_t \nabla_w R_S(w_t)$
η_t ↓ learning rate

Stochastic gradient descent

SGD update} : $w_{t+1} = w_t - \eta_t \hat{\nabla}_w R_S(w_t)$

(noisy estimate of true gradient)

$$\hat{\nabla}_w R_S(w_t) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla_w \ell((x_i, y_i), h(x_i, w))$$

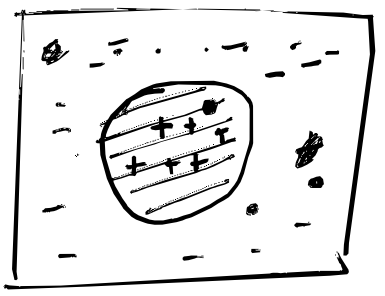
Workhorse
AdamW (momentum)
accelerated version of GD/SGD.

If you solve ERM, do you have guarantees on the generalization error?

Ben-David, Shalev-Schwartz

Example:

$\mathcal{X}$



$\mathcal{D} \equiv$ Unif.

$$Ar(\,\cdot\,) = \frac{Ar(\mathcal{X})}{2}$$

Target function: $f(x) = \begin{cases} 1 & x \in \\ -1 & \text{o.w.} \end{cases}$

$(x_i, y_i) \sim \mathcal{D}$. $\quad S = \{(x_i, y_i)\}_{i=1}^{m} \sim \mathcal{D}^m$

training sample

$$h(x) = \begin{cases} y_i & x = x_i \\ 1 & \text{o.w.} \end{cases}$$

Empirical risk $\boxed{R_S(h) = 0}$.

Generalization error

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}\, \ell((x,y), h) = \,? = 1/2.$$

Overfitting / memorization:

$R_S(h) = 0$ but gen. error is high.

$$R_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell((x_i, y_i), h)$$

$$= \frac{1}{m}\sum_{i=1}^{m} 0 = 0.$$