PAC-Bayes, Sample Compress & Kernel Methods

Pascal Germain

Joint work with François Laviolette, Alexandre Lacasse,
Alexandre Lacoste, Mario Marchand and Sara Shanian

GRAAL
(Université Laval, Québec city)

March 22, 2010

Outline

In this lecture, we will :

- Review some elements of the **Sample-Compress theory**
- See how we can describe a SVM as a **Majority Vote of Sample-Compressed classifiers** (the Sc-SVM)
- Use the **PAC-Bayes** theory to **upper-bound** the risk of our Sc-SVM
- **Minimize** this PAC-Bayes bound and present **experimental results**
- and Conclude...

Outline

In this lecture, we will :

- Review some elements of the **Sample-Compress theory**
- See how we can describe a SVM as a **Majority Vote of Sample-Compressed classifiers** (the Sc-SVM)
- Use the **PAC-Bayes** theory to **upper-bound** the risk of our Sc-SVM
- **Minimize** this PAC-Bayes bound and present **experimental results**
- and Conclude...

The Classification problem

We consider a training set S of m examples

$$S \overset{\text{def}}{=} (\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m)$$

where each $\mathbf{z}_i$ is a input-output pair:

$$\begin{array}{lcll} \mathbf{z}_i & \overset{\text{def}}{=} & (\mathbf{x}_i, y_i) & \\ \mathbf{x}_i & \in & \mathcal{X} \subseteq \mathbb{R}^n & \textit{(Real atttibutes)} \\ y_i & \in & \mathcal{Y} = \{-1, +1\} & \textit{(Binary classif.)} \end{array}$$

Each example $\mathbf{z}_i$ is drawn *IID* according to an unknown probability distribution D on $\mathcal{X} \times \mathcal{Y}$. Hence :

$$S \sim D^m$$

.

Elements of the Sample Compression theory

A **sc-classifier** $h_{\mathbf{i}}^{\mu}$ is a data-dependent classifier described by two variables:

- A **compression-set** $S_{\mathbf{i}}$ containing a subset of the training sequence S describing the classifier
  - $\mathbf{i} \overset{\text{def}}{=} \langle i_1, i_2, \ldots, i_{|\mathbf{i}|} \rangle \quad \text{with} \quad 1 \leq i_1 < i_2 < \ldots < i_{|\mathbf{i}|} \leq m$
- A **message string** μ containing the additional information needed to construct the classifier.
  - μ is choosen among $\mathcal{M}_{\mathbf{i}}$, a predefined set of all messages that can be supplied with $S_{\mathbf{i}}$.

Given $S_{\mathbf{i}}$ and μ, a **reconstruction function** $\mathcal{R}$ outputs a classifier :

$$h_{\mathbf{i}}^{\mu} \overset{\text{def}}{=} \mathcal{R}(S_{\mathbf{i}}, \mu)\,.$$

Risk of a sc-classifier

The **risk** (or generalization error) of a classifier h is defined as

$$R_D(h) \overset{\text{def}}{=} \mathop{\mathbf{E}}_{(\mathbf{x},y)\sim D} I(h(\mathbf{x}) \neq y) = \Pr_{(\mathbf{x},y)\sim D} (h(\mathbf{x}) \neq y)$$

where $I(a) = 1$ if predicate a is true and 0 otherwise.

The **empirical risk** of a sc-classifier $h_{\mathbf{i}}^{\mu}$ on the training set S is defined by

$$R_S(h_{\mathbf{i}}^{\mu}) \overset{\text{def}}{=} \frac{1}{m} \sum_{j=1}^{m} R_{\langle(\mathbf{x}_j,y_j)\rangle}(h_{\mathbf{i}}^{\mu}),$$

where

$$R_{\langle(\mathbf{x}_j,y_j)\rangle}(h_{\mathbf{i}}^{\mu}) \overset{\text{def}}{=} \begin{cases} I(h_{\mathbf{i}}^{\mu}(\mathbf{x}_j) \neq y_j) & \text{if } j \notin \mathbf{i} \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $mR_s(h_{\mathbf{i}}^{\mu}) \sim \mathrm{Bin}\Big(m - \|\mathbf{i}\|, R_D(h_{\mathbf{i}}^{\mu})\Big)$.

Outline

In this lecture, we will :

- Review some elements of the **Sample-Compress theory**
- See how we can describe a SVM as a **Majority Vote of Sample-Compressed classifiers** (the Sc-SVM)
- Use the **PAC-Bayes** theory to **upper-bound** the risk of our Sc-SVM
- **Minimize** this PAC-Bayes bound and present **experimental results**
- and Conclude...

Redefining the SVM as a Majority Vote of sc-classifiers

We denote $\mathcal{H}^S$ the set of all sc-classifiers. Each $h_{\mathbf{i}}^{\mu} \in \mathcal{H}^S$ is such as :

- The **compression-set** contains one training example :

$$S_{\mathbf{i}} \in \{S_{\langle 1 \rangle}, S_{\langle 2 \rangle}, \ldots, S_{\langle m \rangle}\}$$

- The **message string** is formed by a real number and a sign :

$$\mu \in \mathcal{M}_{\mathbf{i}} = [-1, 1] \times \{+, -\}$$

We consider **pairs of boolean complement classifiers** such as :

$$h_{\mathbf{i}}^{(\sigma,-)}(\mathbf{x}) = -h_{\mathbf{i}}^{(\sigma,+)}(\mathbf{x}) \quad \forall\, \mathbf{x} \in \mathcal{X},\, \sigma \in [-1, 1]\,.$$

sc-classifier $h_{\mathbf{i}}^{\mu} \in \mathcal{H}^S$

Comp-set: $S_{\mathbf{i}} \in \{S_{\langle 1 \rangle}, \ldots, S_{\langle m \rangle}\}$
Message: $\mu \in \mathcal{M}_{\mathbf{i}} = [-1,1] \times \{+,-\}$

Distribution Q

$Q(h_{\mathbf{i}}^{\mu}) = Q_{\mathcal{I}}(\mathbf{i}) Q_{S_{\mathbf{i}}}(\mu)$
$Q(h_{\mathbf{i}}^{(\sigma,+)}) - Q(h_{\mathbf{i}}^{(\sigma,-)}) = w_{\mathbf{i}}$

Let Q be a **probability distribution** over $\mathcal{H}^S$. We denote

- $Q_{\mathcal{I}}$, the probability that a compression-set $S_{\mathbf{i}}$ is chosen by Q:

$$Q_{\mathcal{I}}(\mathbf{i}) \overset{\text{def}}{=} \int_{\mu \in \mathcal{M}_{\mathbf{i}}} Q(h_{\mathbf{i}}^{\mu}) d\mu$$

- $Q_{S_{\mathbf{i}}}$, the probability of choosing message μ given $S_{\mathbf{i}}$:

$$Q_{S_{\mathbf{i}}}(\mu) \overset{\text{def}}{=} Q(h_{\mathbf{i}}^{\mu} | S_{\mathbf{i}})$$

- Therefore, $Q(h_{\mathbf{i}}^{\mu}) = Q_{\mathcal{I}}(\mathbf{i}) Q_{S_{\mathbf{i}}}(\mu)$.

The **output** of the majority vote classifier (*bayes classifier*) is given by :

$$B_Q(\mathbf{x}) \overset{\text{def}}{=} \operatorname{sgn}\left[\mathop{\mathbf{E}}_{h \sim Q} h(\mathbf{x}) \right]$$

sc-classifier $h_{\mathbf{i}}^{\mu} \in \mathcal{H}^S$

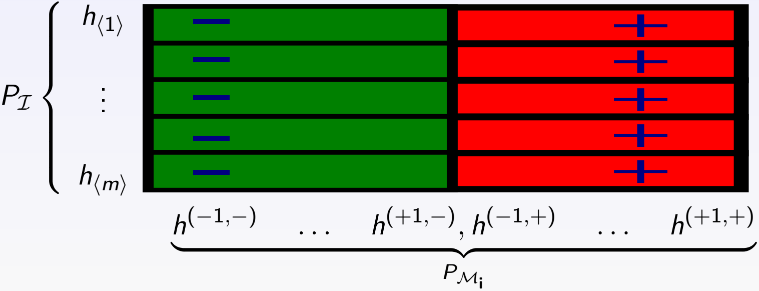
Comp-set: $S_{\mathbf{i}} \in \{S_{\langle 1\rangle}, \ldots, S_{\langle m\rangle}\}$
Message: $\mu \in \mathcal{M}_{\mathbf{i}} = [-1,1] \times \{+,-\}$

Distribution Q

$Q(h_{\mathbf{i}}^{\mu}) = Q_{\mathcal{I}}(\mathbf{i}) Q_{S_{\mathbf{i}}}(\mu)$
$Q(h_{\mathbf{i}}^{(\sigma,+)}) - Q(h_{\mathbf{i}}^{(\sigma,-)}) = w_{\mathbf{i}}$

Before seing the data, we define a **prior distribution** over the compression-sets and the message strings. This gives us indirectly a prior P over $\mathcal{H}^S$ such as :

- $P_{\mathcal{I}}$ is an uniform distribution over all possible compression-sets ;
- For each compression-set $S_{\mathbf{i}}$, $P_{S_{\mathbf{i}}}$ is uniform over all messages.



$P_{\mathcal{I}}$
$h_{\langle 1\rangle}$
$\vdots$
$h_{\langle m\rangle}$
$-$
$+$
$h^{(-1,-)}$ $\ldots$ $h^{(+1,-)}, h^{(-1,+)}$ $\ldots$ $h^{(+1,+)}$
$P_{\mathcal{M}_{\mathbf{i}}}$

sc-classifier $h_{\mathbf{i}}^{\mu} \in \mathcal{H}^S$

Comp-set: $S_{\mathbf{i}} \in \{S_{\langle 1 \rangle}, \ldots, S_{\langle m \rangle}\}$
Message: $\mu \in \mathcal{M}_{\mathbf{i}} = [-1,1] \times \{+,-\}$

Distribution Q

$Q(h_{\mathbf{i}}^{\mu}) = Q_{\mathcal{I}}(\mathbf{i}) Q_{S_{\mathbf{i}}}(\mu)$
$Q(h_{\mathbf{i}}^{(\sigma,+)}) - Q(h_{\mathbf{i}}^{(\sigma,-)}) = w_{\mathbf{i}}$

We say that a posterior Q is **aligned on** a prior P when for all $\mathbf{i}$ and σ:

$$Q(h_{\mathbf{i}}^{(\sigma,+)}) + Q(h_{\mathbf{i}}^{(\sigma,-)}) = P(h_{\mathbf{i}}^{(\sigma,+)}) + P(h_{\mathbf{i}}^{(\sigma,-)})$$

Moreover, we say that a posterior Q is **strongly aligned** when for all $\mathbf{i}$, there is a $w_{\mathbf{i}}$ such that for all σ:

$$Q(h_{\mathbf{i}}^{(\sigma,+)}) - Q(h_{\mathbf{i}}^{(\sigma,-)}) = w_{\mathbf{i}}$$

By restricting ourself to strongly aligned posterior, we obtain a posterior distribution totally defined by the $w_{\mathbf{i}}$'s :

$$\begin{aligned} Q(h_{\mathbf{i}}^{(\sigma,+)}) &= \frac{1}{2}\left(P(h_{\mathbf{i}}^{(\sigma,+)}) + P(h_{\mathbf{i}}^{(\sigma,-)}) + w_{\mathbf{i}}\right) \\ Q(h_{\mathbf{i}}^{(\sigma,-)}) &= \frac{1}{2}\left(P(h_{\mathbf{i}}^{(\sigma,+)}) + P(h_{\mathbf{i}}^{(\sigma,-)}) - w_{\mathbf{i}}\right) \end{aligned}$$

sc-classifier $h_{\mathbf{i}}^{\mu} \in \mathcal{H}^{S}$

Comp-set: $S_{\mathbf{i}} \in \{S_{\langle 1\rangle}, \ldots, S_{\langle m\rangle}\}$
Message: $\mu \in \mathcal{M}_{\mathbf{i}} = [-1,1] \times \{+,-\}$

Distribution Q

$Q(h_{\mathbf{i}}^{\mu}) = Q_{\mathcal{I}}(\mathbf{i}) Q_{S_{\mathbf{i}}}(\mu)$
$Q(h_{\mathbf{i}}^{(\sigma,+)}) - Q(h_{\mathbf{i}}^{(\sigma,-)}) = w_{\mathbf{i}}$

We say that a posterior Q is **aligned on** a prior P when for all $\mathbf{i}$ and σ:

$$Q(h_{\mathbf{i}}^{(\sigma,+)}) + Q(h_{\mathbf{i}}^{(\sigma,-)}) = P(h_{\mathbf{i}}^{(\sigma,+)}) + P(h_{\mathbf{i}}^{(\sigma,-)})$$

Moreover, we say that a posterior Q is **strongly aligned** when for all $\mathbf{i}$, there is a $w_{\mathbf{i}}$ such that for all σ:

$$Q(h_{\mathbf{i}}^{(\sigma,+)}) - Q(h_{\mathbf{i}}^{(\sigma,-)}) = w_{\mathbf{i}}$$

By restricting ourself to strongly aligned posterior, we obtain a posterior distribution totally defined by the $w_{\mathbf{i}}$'s :

$$\begin{aligned} Q(h_{\mathbf{i}}^{(\sigma,+)}) &= \frac{1}{2}\left(P(h_{\mathbf{i}}^{(\sigma,+)}) + P(h_{\mathbf{i}}^{(\sigma,-)}) + w_{\mathbf{i}}\right) \\ Q(h_{\mathbf{i}}^{(\sigma,-)}) &= \frac{1}{2}\left(P(h_{\mathbf{i}}^{(\sigma,+)}) + P(h_{\mathbf{i}}^{(\sigma,-)}) - w_{\mathbf{i}}\right) \end{aligned}$$

sc-classifier $h_{\mathbf{i}}^{\mu} \in \mathcal{H}^{S}$

Comp-set: $S_{\mathbf{i}} \in \{S_{\langle 1\rangle}, \ldots, S_{\langle m\rangle}\}$

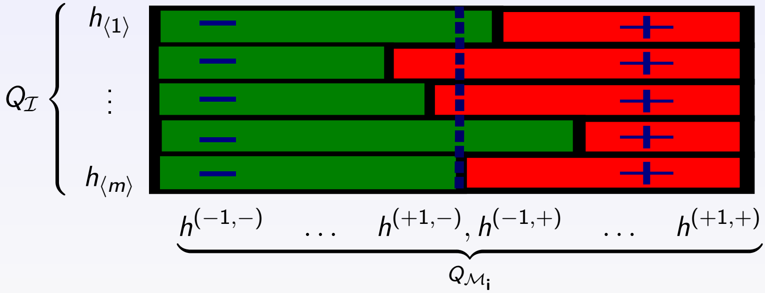
Message: $\mu \in \mathcal{M}_{\mathbf{i}} = [-1,1] \times \{+,-\}$

Distribution Q

$Q(h_{\mathbf{i}}^{\mu}) = Q_{\mathcal{I}}(\mathbf{i})\, Q_{S_{\mathbf{i}}}(\mu)$

$Q(h_{\mathbf{i}}^{(\sigma,+)}) - Q(h_{\mathbf{i}}^{(\sigma,-)}) = w_{\mathbf{i}}$

$$
\begin{aligned}
Q(h_{\mathbf{i}}^{(\sigma,+)}) &= \frac{1}{2}\left(P(h_{\mathbf{i}}^{(\sigma,+)}) + P(h_{\mathbf{i}}^{(\sigma,-)}) + w_{\mathbf{i}}\right)\\
Q(h_{\mathbf{i}}^{(\sigma,-)}) &= \frac{1}{2}\left(P(h_{\mathbf{i}}^{(\sigma,+)}) + P(h_{\mathbf{i}}^{(\sigma,-)}) - w_{\mathbf{i}}\right)
\end{aligned}
$$



$Q_{\mathcal{I}}$
$h_{\langle 1\rangle}$
⋮
$h_{\langle m\rangle}$
−
+
$h^{(-1,-)}$ … $h^{(+1,-)}, h^{(-1,+)}$ … $h^{(+1,+)}$
$Q_{\mathcal{M}_{\mathbf{i}}}$

sc-classifier $h_{\mathbf{i}}^{\mu} \in \mathcal{H}^S$

Comp-set: $S_{\mathbf{i}} \in \{S_{\langle 1\rangle}, \ldots, S_{\langle m\rangle}\}$
Message: $\mu \in \mathcal{M}_{\mathbf{i}} = [-1,1] \times \{+,-\}$

Distribution Q

$Q(h_{\mathbf{i}}^{\mu}) = Q_{\mathcal{I}}(\mathbf{i}) Q_{S_{\mathbf{i}}}(\mu)$
$Q(h_{\mathbf{i}}^{(\sigma,+)}) - Q(h_{\mathbf{i}}^{(\sigma,-)}) = w_{\mathbf{i}}$

Consider any similarity function $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to [-1,1]$.
We say that **reconstruction function** $\mathcal{R}$ is associated to k when :

$$h_{\langle i\rangle}^{(\sigma,+)}(\mathbf{x}) \overset{\text{def}}{=} \begin{cases} +1 & \text{if } \sigma < k(\mathbf{x}_i, \mathbf{x}) \\ -1 & \text{otherwise} \end{cases}$$

$$h_{\mathbf{i}}^{(\sigma,-)}(\mathbf{x}) \overset{\text{def}}{=} -h_{\mathbf{i}}^{(\sigma,+)}(\mathbf{x})\,.$$

We finally obtain that our strongly aligned posterior will be such that:

$$Q_{\mathcal{I}}(\langle i\rangle) = \frac{1}{m}\,, \quad w_{\langle i\rangle} \cdot k(\mathbf{x}_i, \mathbf{x}) = \int_{\mu \in \mathcal{M}_{\langle i\rangle}} h_{\langle i\rangle}^{\mu}(\mathbf{x}) \cdot Q_{\langle i\rangle}(\mu)\; d\mu\,.$$

Thus, the majority vote output $B_Q(\mathbf{x}) = \operatorname{sgn}\left[\mathop{\mathbf{E}}_{h\sim Q} h(\mathbf{x})\right]$ will be the same as
$f_{\text{SVM}}(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i k(\mathbf{x}_i, \mathbf{x})\right)$ when $w_{\langle i\rangle} = \dfrac{y_i \alpha_i}{Z \cdot m}$. $\left(Z \overset{\text{def}}{=} \sum_{i=1}^{m} \alpha_i\right)$

sc-classifier $h_{\mathbf{i}}^{\mu} \in \mathcal{H}^S$

Comp-set: $S_{\mathbf{i}} \in \{S_{\langle 1 \rangle}, \ldots, S_{\langle m \rangle}\}$
Message: $\mu \in \mathcal{M}_{\mathbf{i}} = [-1,1] \times \{+,-\}$

Distribution Q

$Q(h_{\mathbf{i}}^{\mu}) = Q_{\mathcal{I}}(\mathbf{i}) Q_{S_{\mathbf{i}}}(\mu)$
$Q(h_{\mathbf{i}}^{(\sigma,+)}) - Q(h_{\mathbf{i}}^{(\sigma,-)}) = w_{\mathbf{i}}$

Consider any similarity function $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \rightarrow [-1,1]$.
We say that **reconstruction function** $\mathcal{R}$ is associated to k when :

$$h_{\langle i \rangle}^{(\sigma,+)}(\mathbf{x}) \overset{\text{def}}{=} \begin{cases} +1 & \text{if } \sigma < k(\mathbf{x}_i, \mathbf{x}) \\ -1 & \text{otherwise} \end{cases}$$

$$h_{\mathbf{i}}^{(\sigma,-)}(\mathbf{x}) \overset{\text{def}}{=} -h_{\mathbf{i}}^{(\sigma,+)}(\mathbf{x}) .$$

We finally obtain that our strongly aligned posterior will be such that:

$$Q_{\mathcal{I}}(\langle i \rangle) = \frac{1}{m}, \quad w_{\langle i \rangle} \cdot k(\mathbf{x}_i, \mathbf{x}) = \int_{\mu \in \mathcal{M}_{\langle i \rangle}} h_{\langle i \rangle}^{\mu}(\mathbf{x}) \cdot Q_{\langle i \rangle}(\mu) \, d\mu .$$

Thus, the majority vote output $B_Q(\mathbf{x}) = \operatorname{sgn}\left[\mathop{\mathbf{E}}_{h\sim Q} h(\mathbf{x})\right]$ will be the same as $f_{\text{SVM}}(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i\alpha_i k(\mathbf{x}_i, \mathbf{x})\right)$ when $w_{\langle i \rangle} = \dfrac{y_i\alpha_i}{Z \cdot m}$. $\left(Z \overset{\text{def}}{=} \sum_{i=1}^{m} \alpha_i\right)$

Outline

In this lecture, we will :

- Review some elements of the **Sample-Compress theory**
- See how we can describe a SVM as a **Majority Vote of Sample-Compressed classifiers** (the Sc-SVM)
- Use the **PAC-Bayes** theory to **upper-bound** the risk of our Sc-SVM
- **Minimize** this PAC-Bayes bound and present **experimental results**
- and Conclude...

PAC-Bayes bounds for Sc-SVM

PAC-Bayes theorems allow us to bound the risk of a majority vote classifier B_Q by bounding the **risk of the Gibbs classifier** G_Q, related to B_Q

- Given $\mathbf{x}$, G_Q draws h according to Q and classifies $\mathbf{x}$ according to h.
- It follows that $R_D(B_Q) \leq 2R_D(G_Q)$.

In our setting, the Gibbs risk $R_D(G_Q)$ will be likely near $1/2$, even if the Bayes risk is close to 0.

- Each sc-classifier $h_{\mathbf{i}}^{\mu} \in \mathcal{H}^S$ might be really weak.

We want to bound a **more relevant risk**!

Similary at [Germain et al. *PAC-Bayes bounds for general loss functions* (2006)], we can consider any non-negative loss ζ that can be expended by a Taylor series around the margin $M_Q(\mathbf{x}, y) = 0$.

PAC-Bayes bounds for Sc-SVM

PAC-Bayes theorems allow us to bound the risk of a majority vote classifier B_Q by bounding the **risk of the Gibbs classifier** G_Q, related to B_Q

- Given $\mathbf{x}$, G_Q draws h according to Q and classifies $\mathbf{x}$ according to h.
- It follows that $R_D(B_Q) \leq 2R_D(G_Q)$.

In our setting, the Gibbs risk $R_D(G_Q)$ will be likely near $1/2$, even if the Bayes risk is close to 0.

- Each sc-classifier $h_{\mathbf{i}}^{\mu} \in \mathcal{H}^S$ might be really weak.

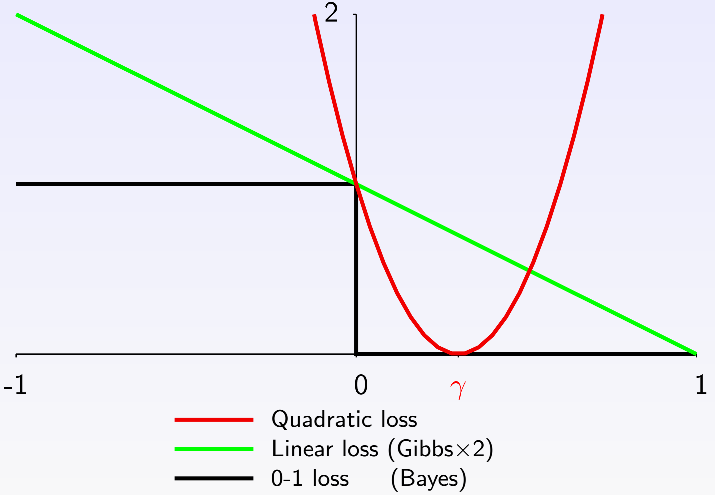
We want to bound a **more relevant risk**!

Similary at [Germain et al. *PAC-Bayes bounds for general loss functions* (2006)], we can consider any non-negative loss ζ that can be expended by a Taylor series around the margin $M_Q(\mathbf{x}, y) = 0$.

Margin of the majority vote classifier

$$M_Q(\mathbf{x}, y) \quad \overset{\text{def}}{=} \quad \mathbf{E}_{h_{\mathbf{i}}^{\mu}\sim Q}\; y\, h_{\mathbf{i}}^{\mu}(\mathbf{x}) \quad = \quad 1-2R_{\langle(\mathbf{x}_j,y_j)\rangle}(G_Q)$$

We choose to use the **quadratic loss** function $\zeta_\gamma(\alpha) = \left(1 - \frac{1}{\gamma}\alpha\right)^2$.



First PAC-Bayes theorem

We adapted the **Catoni's theorem** to consider:

- A general loss function ζ
- A set of (data-dependent) sc-classifiers of size $\leq l$

Theorem

For any D, any family $(\mathcal{H}^S)_{S\in\mathcal{D}^m}$ of sets of sc-classifiers of size at most l, any prior P, any $\delta \in (0,1]$, any $C_1 \in \mathbb{R}^+$ and any margin loss function ζ of degree $\frac{m}{l}$:

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall\, Q \text{ on } \mathcal{H}^S\!: \\ \zeta_D^Q \quad \leq \quad C' \cdot \left(\zeta_S^Q + \dfrac{\zeta'(1)\cdot \mathrm{KL}(Q\|P) + \zeta(1)\cdot \ln\frac{1}{\delta}}{C_1\cdot m}\right)\end{array}\right) \geq 1-\delta\,,$$

where $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence and $C' = \dfrac{C_1\cdot\frac{m}{m-l\cdot\deg\zeta}}{1-e^{-C_1\cdot\frac{m-l\cdot\deg\zeta}{m}}}$.

Finding Q that minimizes this bound is equivalent to finding Q minimizing:

$$f(Q) \overset{\mathrm{def}}{=} C\cdot\zeta_S^Q + KL(Q\|P) \qquad \text{(where } C \text{ is an hyperparameter)}$$

Second PAC-Bayes theorem

We adapted the **Langford and Seeger's theorem** which use the KL divergence between two Bernoulli distributions of prob of success p and q:

$$\mathrm{kl}(q\|p) \quad \overset{\mathrm{def}}{=} \quad q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p} \quad = \quad \mathrm{kl}(1-q\|1-p)$$

The usual term $\mathrm{KL}(Q\|P)$ disappear as we consider aligned posteriors:

$$Q(h) + Q(-h) = P(h) + P(-h)\ \forall h \in \mathcal{H}$$

Theorem

For any D, any family $(\mathcal{H}^S)_{S\in\mathcal{D}^m}$ of sets of sc-classifiers of size at most l, any prior P, any $\delta \in (0,1]$, any margin loss function ζ of degree $< m/l$, we have

$$\Pr_{S\sim D^m}\left(\begin{array}{l}\forall Q\in\mathcal{H}^S \ \textbf{aligned on } P: \\ \mathrm{kl}\left(\frac{1}{\zeta(1)}\cdot\zeta_S^Q \middle\| \frac{1}{\zeta(1)}\cdot\zeta_D^Q\right) \le \frac{\ln\frac{m+1}{\delta}}{m-l\cdot\deg\zeta}\end{array}\right) \ge 1-\delta$$

This bound suggests to minimize the empirical risk: $f(Q) \overset{\mathrm{def}}{=} \zeta_S^Q$

We want to bound random variable $\mathop{\mathbf{E}}_{h\sim P} e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))}$ in term of $R(G_Q)$.

General theorem

Term $\mathrm{KL}(Q\|P)$ arises when transforming expectation over P into expectation over Q:

$$
\begin{aligned}
& \ln\left[\mathop{\mathbf{E}}_{h\sim P} e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))}\right] \\
= \ & \ln\left[\mathop{\mathbf{E}}_{h\sim Q} \frac{P(h)}{Q(h)} e^{m\cdot \mathrm{kl}(R_S(h),R(h))}\right] \\
\geq \ & \mathop{\mathbf{E}}_{h\sim Q} \ln\left[\frac{P(h)}{Q(h)} e^{m\cdot \mathrm{kl}(R_S(h),R(h))}\right] \\
= \ & m \mathop{\mathbf{E}}_{h\sim Q} \mathrm{kl}(R_S(h),R(h)) - \mathrm{KL}(Q\|P) \\
\geq \ & m\cdot \mathrm{kl}(\mathop{\mathbf{E}}_{h\sim Q} R_S(h), \mathop{\mathbf{E}}_{h\sim Q} R(h)) - \mathrm{KL}(Q\|P) \\
= \ & m\cdot \mathrm{kl}(R_S(G_Q),R(G_Q)) - \mathrm{KL}(Q\|P)\ .
\end{aligned}
$$

Aligned posterior theorem

Here, we do the same operation for "free" (proof on next slide):

$$
\begin{aligned}
& \ln\left[\mathop{\mathbf{E}}_{h\sim P} e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))}\right] \\
= \ & \ln\left[\mathop{\mathbf{E}}_{h\sim Q} e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))}\right] \\
\geq \ & \mathop{\mathbf{E}}_{h\sim Q} \ln\left[e^{m\cdot \mathrm{kl}(R_S(h),R(h))}\right] \\
= \ & m \mathop{\mathbf{E}}_{h\sim Q} \mathrm{kl}(R_S(h),R(h)) \\
\geq \ & m\cdot \mathrm{kl}(\mathop{\mathbf{E}}_{h\sim Q} R_S(h), \mathop{\mathbf{E}}_{h\sim Q} R(h)) \\
= \ & m\cdot \mathrm{kl}(R_S(G_Q),R(G_Q))\ .
\end{aligned}
$$

The two "$\geq$" come from Jensen's inequality: $\mathbf{E}\, f(X) \geq f(\mathbf{E}\, X)$ for convex f.

First, note that as we have $h \in \mathcal{H}^S \Rightarrow -h \in \mathcal{H}^S$:

$$\mathop{\mathbf{E}}_{h\sim P} e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))} = \int_{h\in\mathcal{H}} P(h) e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))} = \int_{h\in\mathcal{H}} P(-h) e^{m\cdot \mathrm{kl}(R_S(-h)\|R(-h))} .$$

Then, following that $Q(h) + Q(-h) = P(h) + P(-h)$:

$$\begin{aligned}
& 2 \mathop{\mathbf{E}}_{h\sim P} e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))} \\
&= \int_{h\in\mathcal{H}} P(h) e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))} + \int_{h\in\mathcal{H}} P(-h) e^{m\cdot \mathrm{kl}(R_S(-h)\|R(-h))} \\
&= \int_{h\in\mathcal{H}} P(h) e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))} + \int_{h\in\mathcal{H}} P(-h) e^{m\cdot \mathrm{kl}(1-R_S(h)\|1-R(h))} \\
&= \int_{h\in\mathcal{H}} \left(P(h) + P(-h)\right) e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))} \\
&= \int_{h\in\mathcal{H}} \left(Q(h) + Q(-h)\right) e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))} \\
&= \int_{h\in\mathcal{H}} Q(h) e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))} + \int_{h\in\mathcal{H}} Q(-h) e^{m\cdot \mathrm{kl}(R_S(-h)\|R(-h))} \\
&= 2 \mathop{\mathbf{E}}_{h\sim Q} e^{m\cdot \mathrm{kl}(R_S(h)\|R(h))} .
\end{aligned}$$

Outline

In this lecture, we will :

- Review some elements of the **Sample-Compress theory**
- See how we can describe a SVM as a **Majority Vote of Sample-Compressed classifiers** (the Sc-SVM)
- Use the **PAC-Bayes** theory to **upper-bound** the risk of our Sc-SVM
- **Minimize** this PAC-Bayes bound and present **experimental results**
- and Conclude...

We have designed two learning algorithms

Remember that Q is **strongly aligned**: $Q(h_{\mathbf{i}}^{(\sigma,+)}) - Q(h_{\mathbf{i}}^{(\sigma,-)}) = w_{\mathbf{i}}$.

The task of the algorithms is to find a vector $\mathbf{w} = (w_1, \ldots, w_m)$,

$$\begin{aligned} w_i &\overset{\text{def}}{=} w_{\langle i \rangle} = Q(h_{\langle i \rangle}^{(\sigma,+)}) - Q(h_{\langle i \rangle}^{(\sigma,-)}) \\ |w_j| &\leq \frac{1}{m} \quad \forall j \in \{1, \ldots, m\} \end{aligned}$$

The empirical margin $\widehat{M}_Q$ is now defined by

$$\widehat{M}_Q(\mathbf{x}_j, y_j) = \sum_{k=0}^{m} y_j\, w_k\, \widehat{G}(\mathbf{x}_k, \mathbf{x}_j)$$

where

$$\widehat{G}(\mathbf{x}_j, \mathbf{x}_l) \overset{\text{def}}{=} \begin{cases} k(\mathbf{x}_j, \mathbf{x}_l) & \forall j \in \{1, .., m\} \text{ and } j \neq l \\ 1 & \forall j \in \{1, .., m\} \text{ and } j = l \end{cases}$$

We have designed two learning algorithms

Remember that Q is **strongly aligned**: $Q(h_{\mathbf{i}}^{(\sigma,+)}) - Q(h_{\mathbf{i}}^{(\sigma,-)}) = w_{\mathbf{i}}$.

The task of the algorithms is to find a vector $\mathbf{w} = (w_1, \ldots, w_m)$,

$$\begin{aligned} w_i &\overset{\text{def}}{=} w_{\langle i\rangle} = Q(h_{\langle i\rangle}^{(\sigma,+)}) - Q(h_{\langle i\rangle}^{(\sigma,-)}) \\ |w_j| &\leq \frac{1}{m} \quad \forall j \in \{1, \ldots, m\} \end{aligned}$$

The empirical margin $\widehat{M}_Q$ is now defined by

$$\widehat{M}_Q(\mathbf{x}_j, y_j) = \sum_{k=0}^{m} y_j\, w_k\, \widehat{G}(\mathbf{x}_k, \mathbf{x}_j)$$

where

$$\widehat{G}(\mathbf{x}_j, \mathbf{x}_l) \overset{\text{def}}{=} \begin{cases} k(\mathbf{x}_j, \mathbf{x}_l) & \forall\, j \in \{1, .., m\} \text{ and } j \neq l \\ 1 & \forall\, j \in \{1, .., m\} \text{ and } j = l \end{cases}$$

Algorithm with *KL* (Based on our first PAC-Bayes theorem)

Find $\mathbf{w}$ that minimizes $f(\mathbf{w}) \stackrel{\mathrm{def}}{=} C \cdot \sum_{j=0}^{m} \zeta_\gamma \left(y_j \, \mathbf{w} \, \widehat{\mathbf{G}}(\mathbf{x}_j) \right) + \mathrm{KL}(Q_{\mathbf{w}} || P)$

Parameters to tune :

- C, the trade-off between the two terms to minimize
- γ, the minimum of the quadratic risk

Algorithm without *KL* (Based on our second PAC-Bayes theorem)

Find $\mathbf{w}$ that minimizes $f(\mathbf{w}) \stackrel{\mathrm{def}}{=} \sum_{j=0}^{m} \zeta_\gamma \left(y_j \, \mathbf{w} \, \widehat{\mathbf{G}}(\mathbf{x}_j) \right)$

Parameter to tune :

- γ, the minimum of the quadratic risk

Both objective functions are **convex**. ⇒ Only one global minimum.

Experimental results (RBF kernel, 10-folds CV)

Dataset	$\|T\|$	$\|S\|$	n	**Classic SVM**	**SC-SVM** (with KL)	**SC-SVM** (w/o KL)
Usvotes	200	235	16	0.065	**0.060**	**0.060**
Liver	175	170	6	**0.303**	0.371	**0.303**
Credit-A	300	353	15	0.187	0.170	**0.150**
Glass	107	107	9	0.159	**0.131**	0.178
Haberman	150	144	3	**0.273**	0.287	0.287
Heart	147	150	13	0.184	**0.163**	0.190
sonar	104	104	60	0.183	0.144	**0.135**
BreastCancer	340	343	9	0.038	**0.035**	**0.035**
Tic-tac-toe	479	479	9	0.023	**0.015**	**0.015**
Ionosphere	175	176	34	0.051	**0.029**	**0.029**
Wdbc	284	285	30	0.070	0.092	**0.067**
MNIST:0vs8	1916	500	784	0.005	**0.004**	**0.004**
MNIST:1vs7	1922	500	784	0.012	**0.008**	0.010
MNIST:1vs8	1936	500	784	0.013	**0.011**	**0.011**
MNIST:2vs3	1905	500	784	0.023	**0.016**	0.018
Letter:AB	1055	500	16	**0.001**	**0.001**	**0.001**
Letter:DO	1058	500	16	0.013	**0.009**	**0.009**
Letter:OQ	1036	500	16	**0.014**	0.017	0.017
Adult	10000	1809	14	0.160	**0.157**	**0.157**
Mushroom	4062	4062	22	**0.000**	**0.000**	**0.000**
Waveform	4000	4000	21	**0.068**	0.069	**0.068**
Ringnorm	3700	3700	20	0.015	0.016	**0.012**

Outline

In this lecture, we will :

- Review some elements of the **Sample-Compress theory**
- See how we can describe a SVM as a **Majority Vote of Sample-Compressed classifiers** (the Sc-SVM)
- Use the **PAC-Bayes** theory to **upper-bound** the risk of our Sc-SVM
- **Minimize** this PAC-Bayes bound and present **experimental results**
- and Conclude...

Future works

We presented a general framework to apply the PAC-Bayes theory to kernels methods.

For now, we compare ourself to the Support Vector Machine, but there is many other possibilities.

Three future research ideas (among others) :

- Experimentations with **non-PSD kernels**
- Consider a majority vote of sc-classifiers of **maximum size** > 1
- Consider **non-strongly aligned** posteriors.