Lecture 1 - setup
Tuesday, January 9, 2024  9:42 AM


### Learning problem

Given a training dataset $D = (x_i, y_i)_{i=1}^n$,

good: learn a prediction mapping $f_w : X \rightarrow \mathcal{Y}$

That has low classification error

generalization error

statistical decision loss

[see lecture 5 of my POM on review of statistical]

"Risk" in ML

Examples of learning approaches:

- Regularized ERM
- Empirical risk minimization

 Empirical error

\[ \hat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, h_w(x_i)) + R(w) \]

Not in $\mathbb{R}$ in $w$

Non-convex

Mussy $\Sigma \text{NP hard to minimize in general}$

\[ \hat{L}(w) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, w) + R(w) \]

\( \Sigma \text{replace } \hat{L} \text{ with a surrogate loss} \)

\( \text{e.g. convex in } w \)

\( \text{ surrogate loss } \)

1 O-lasso

structured sparse line

1 G-lasso
Some important aspects of structured prediction (in binary classification):  
1) \( Y \) output is usually exponentially big (in natural size of input)  
2) Structured error function \( l(y, y') \)  
3) Sometimes constraints on pairs of \( y \)  
   - Consider word alignment example  
   - English words: French words  
   - "like" ⇔ "aime"  
   - "time" ⇔ "temps"  

Note: Need "encoding \( \xi \),"  
\( y = (y_1, \ldots, y_p) \in \{0, 1\}^p \)  
\( \sim \) \( y = (y_1, \ldots) \in \{0, 1\}^p \)
\( y = (y_{ij})_{i,j} \in \mathbb{R}^{n \times n} \)

\( \text{Possible edges (have all pairs)} \)

Matching constraints:

\[
\begin{align*}
\forall i & : \sum_{j} y_{ij} \leq 1 \\
\forall j & : \sum_{i} y_{ij} \leq 1
\end{align*}
\]