Generalization, Overfitting and Regularization

Fabio A. González Ph.D.

Depto. de Ing. de Sistemas e Industrial
Universidad Nacional de Colombia, Bogotá

March 27, 2007

Outline

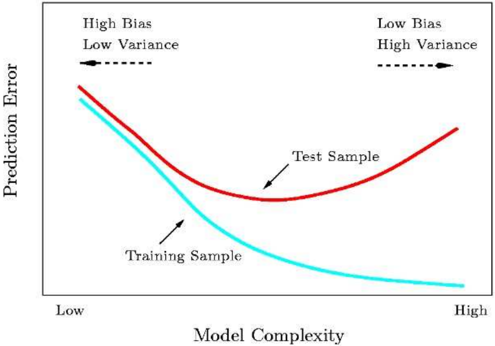
Generalization

- The generalization error is defined as:

$$E[(g(x) - y)^2]$$

- In general, we don't know the exact generalization error of a model, but we can estimate it.
- Alternatives:
    - Training error
    - Test error
    - Mathematical estimation based on model complexity

Training Error and Test Error



Overfitting

- A model overfits if it fits particularities of the training set (noise, bias, etc): low training error - high testing error.
- A complex model has more possibilities to overfit data.
- The generalization error is a function of the model complexity.
- Occam's razor and minimum description length principle.

Dealing with Overfitting

- Break the available data in three subsets:
  - Training
  - Validation
  - Testing
- Train the model varying the complexity
- Use the validation set to estimate the generalization error
- Find the optimal complexity

Bias variance trade-off

- Error could be expressed as:

$$
\begin{aligned}
Error &= variance + Bias^2 \\
&= E[(d - E[d])^2] + (E[d] - \theta)^2
\end{aligned}
$$

- Bias: how much $g(x)$ is wrong
- Variance: how much $g(x)$ fluctuates around expected value

An alternative approach to control overfitting

- Control the complexity in the learning process
- Penalize high-complexity models:

$$L(g(), X) = \text{Prediction Error} + \lambda \text{Complexity}(g())$$

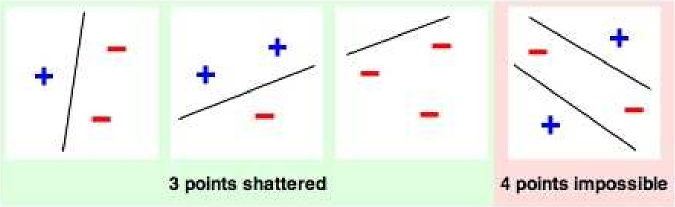
- Complexity can be measured/controlled in different ways

VC Dimension

- Proposed by Vapnik and Chervonenkis:
  V. Vapnik and A. Chervonenkis. "On the uniform convergence of relative frequencies of events to their probabilities." Theory of Probability and its Applications, 16(2):264–280, 1971.
- VC-dimension: Cardinality of the largest set of points that the algorithm can shatter.
- Shattering: A classification model f with some parameter vector Θ is said to shatter a set of data points $(x_1, x_2, \ldots, x_n)$ if, for all assignments of labels to those points, there exists a Θ such that the model f makes no errors when evaluating that set of data points.

Wikipedia:`http://en.wikipedia.org/wiki/VC_dimension`

Example



Rademacher complexity

- Proposed as an alternative to VC dimension:
  Koltchinskii, V. and Panchenko, D. (2000). Rademacher processes and bounding the risk of function learning. In High Dimensional Probability II (E. Giné, D. Mason and J. Wellner, eds.) 443–459. Birkhäuser, Boston.

Definition

For a sample $S = \{\mathrm{x}_1, \dots, \mathrm{x}_l\}$ generated by a distribution $\mathcal{D}$ on a set X and a real-valued function class $\mathcal{F}$ with domain X, the *empirical Rademacher complexity* of $\mathcal{F}$ is the random variable

$$\hat{R}_l(\mathcal{F}) = E_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{l} \sum_{i=1}^{l} \sigma_i f(\mathrm{x}_i) \right| \Bigg| \mathrm{x}_1, \dots, \mathrm{x}_l \right],$$

where $\sigma = \{\sigma_1, \dots, \sigma_l\}$ are independent uniform $\{\pm 1\}$-valued (Rademacher) random variables. The *Rademacher complexity* of $\mathcal{F}$ is

$$R_l(\mathcal{F}) = E_S[\hat{R}_l(\mathcal{F})] = E_{S\sigma} \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{l} \sum_{i=1}^{l} \sigma_i f(\mathrm{x}_i) \right| \right].$$

Bounds on expected error

- **Pattern**: a function $f(x)$ is a pattern in a set of data items generated i.i.d. according to a fixed (but unknown) distribution $\mathcal{D}$ if

$$E_{\mathcal{D}}[f(x)] \approx 0$$

Theorem

Fix $\delta \in (0,1)$ and let $\mathcal{F}$ be a class of functions mapping from Z to $[1, a+1]$. Let $(z_i)_{i=1}^{l}$ be drawn independently according to a probability distribution $\mathcal{D}$. Then with probability at least $1-\delta$ over random draws of samples of size l, every $f \in \mathcal{F}$ satisfies

$$\begin{aligned} E_{\mathcal{D}}[f(z)] &\leq \widehat{E}[f(z)] + R_l(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2l}} \\ &\leq \widehat{E}[f(z)] + \widehat{R}_l(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2l}} \end{aligned}$$