WEIGHTED SUMS OF RANDOM KITCHEN SINKS

Replacing minimization with randomization in learning
The model

- Given a set of training data in a domain
  \[ \{x_i, y_i\}_{i=1...m}, \ x_i \in \mathcal{X}, \ y_i \in \{-1, +1\} \]
- Fit a function to minimize risk
  \[ f : \mathcal{X} \rightarrow \mathbb{R} \]
- Empirical Risk
  \[ R_{emp}[f] \equiv \frac{1}{m} \sum_{i=1}^{m} c(f(x_i, y_i)) \]
- Risk
  \[ R[f] \equiv \mathbb{E}_{(x,y) \sim P} c(f(x, y)) \]
Loss Function

- Hinge loss (SVM)
  \[ \max(0, 1 - yy') \]

- Exponential loss (AdaBoost)
  \[ e^{-yy'} \]

- Quadratic loss
  \[ (y - y')^2 \]
Form of solution function

- Consider solutions in the form

\[ f(x) = \sum_{i=1}^{\infty} \alpha(w_i)\phi(x; w_i) \quad \text{or} \quad f(x) = \int_{\Omega} \alpha(w)\phi(x; w)dw \]

weights feature functions

- Feature functions
  - Eigenfunctions (kernel SVM)
  - Decision trees/stumps (AdaBoost)

\[ \phi(x; w) = \text{sign}(x_{wd} - w_t) \]
\[ \phi(x; w) = \text{sign}(x \cdot w) \]

- More feature functions gives better classification
Solving $f$

- Approximate

$$\min_{w_1, \ldots, w_K, \alpha} \mathbb{R}_\text{emp} \left[ \sum_{k=1}^{K} \phi(x; w_k) \alpha_k \right]$$

- This is hard!

- New approach:
  - Randomly choose $w_k$ and minimize over $\alpha$
Randomized approach

- \( \{x_i, y_i\}_{i=1}^{m} \) Training data
- \(|\phi(x; w)| \leq 1\) Feature function
- \(K\) Number of features
- \(p(w)\) Parameter distribution
- \(C\) Scaling factor

**Algorithm**
- Draw feature parameters randomly from \(p\)
- Let \(z_i \leftarrow [\phi(x_i; w_1), \ldots, \phi(x_i; w_K)]\)
- Minimize empirical risk

\[
\min_{\alpha \in \mathbb{R}^K} \frac{1}{m} \sum_{i=1}^{m} c(\alpha^T z_i, y_i) \quad \text{s.t.} \quad \|\alpha\|_\infty \leq \frac{C}{K}
\]
Experimental Results vs AdaBoost

- Three datasets
  - adult
  - activity
  - KDDCUP99

- Feature function $\phi(x; w) = \text{sign}(x_{wd} - w_t)$
  - $w_d$ sampled uniformly at random
  - $w_t$ sampled from Gaussian
Experimental Results vs AdaBoost
Pros and Cons

• Pros
  • Much faster
  • Allows simple and efficient experimentation of feature functions

• Cons
  • Some loss in quality
  • Need to tune probability distribution (not needed in practice)
Concentration of Risk

- The randomized algorithm returns a function $\hat{f}$ such that

$$R[\hat{f}] - \min_{f \in \mathcal{F}_p} R[f] \leq O\left(\left(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{K}}\right)LC\sqrt{\log \frac{1}{\delta}}\right)$$

with probability $1 - 2\delta$

- $m$ Number of training points
- $K$ Number of feature vectors
- $L$ Lipschitz constant of loss function

- Bound approximation error
  - Lowest risk versus lowest risk from functions returned is not large

- Bound estimation error
  - True risk of every function returned is close to its empirical risk
Proof

- $f^*$ minimizer of risk over all solution functions
- $\hat{f}^*$ minimizer of risk over functions returned
- $\hat{f}$ minimizer of empirical risk over functions returned

\[
R[f^*] \leq R[\hat{f}^*] \leq R[\hat{f}] \underbrace{\text{fixed } w_1, \ldots, w_K}_{\text{fixed } w_1, \ldots, w_K}
\]

- Then

\[
R[\hat{f}] - R[f^*] = R[\hat{f}] - R[\hat{f}^*] + R[\hat{f}^*] - R[f^*] \\
\leq |R[\hat{f}] - R[\hat{f}^*]| + R[\hat{f}^*] - R[f^*] \\
\leq 2\epsilon_{est} + \epsilon_{app}
\]

with probability $1 - 2\delta$
Bounding approximation error

- Lemma 1. Let $\mathbf{X} = \{x_1, \ldots, x_K\}$ be i.i.d. random variables in a ball of radius $M$ centered about the origin in a Hilbert space. Then
  \[ \|\bar{\mathbf{X}} - \mathbb{E}\bar{\mathbf{X}}\| \leq \frac{M}{\sqrt{K}} \left( 1 + \sqrt{2 \log \frac{1}{\delta}} \right) \]
  with probability $1 - \delta$

- Construct functions $f_k = \beta_k \phi(\cdot; w_k)$, $\beta_k = \frac{\alpha(w_k)}{p(w_k)} \Rightarrow \mathbb{E}f_k = f^*$

- Then there exists $\hat{f}(x) = \sum_{k=1}^{K} \frac{\beta_k}{K} \phi(x; w_k)$

- So that
  \[ \sqrt{\int_{\mathcal{X}} \left( \hat{f}(x) - f^*(x) \right)^2 \, d\mu(x)} \leq \frac{C}{\sqrt{K}} \left( 1 + \sqrt{2 \log \frac{1}{d}} \right) \]
Bounding approximation error

- If the loss function has Lipschitz constant $L$, for any two functions

\[
R[f] - R[g] = \mathbb{E}[c(f(x), y) - c(g(x), y)] \leq \mathbb{E}|c(f(x), y) - c(g(x), y)| \\
\leq L\mathbb{E}|f(x) - g(x)| \leq L\sqrt{\mathbb{E}(f(x) - g(x))^2}
\]

- Then

\[
R[\hat{f}] \leq R[f^*] + \frac{LC}{\sqrt{K}} \left(1 + \sqrt{2\log \frac{1}{\delta}}\right)
\]