Adaboost algorithm

- **Choosing \( D_n \)**
  \[ D_{n+1}(x(i)) = D_n(x(i)) \exp \left(-\alpha_n h_n(x(i))d(i)\right)/Z_n \]
  where \( Z_n \) is a normalization constant.

- **Output hypothesis**
  \[ h(x) = \text{sgn} \left( \sum_{n=1:T} \alpha_n h_n(x) \right) \]

- **Last time showed that empirical error decays exponentially as \( T \), the number of times examples are resampled.**

- **Examine generalization error**
Generalization Error using VC bounds

- Using VC bounds can find upper bounds on generalization error
  \[ P(d \neq h(x)) \leq \text{error}(h(S)) + O((Td/m)^{1/2}) \]
  where \(d\) is VC dimension of concept class.
- Error depends on \(T\), the number of times resampling is done, but experimental results indicate that generalization error for boosting decreases as \(T\) increases. Schapire shows how generalization error decreases even when training error is zero.
Generalization error using margin bounds

- Let $\alpha_i$ be a sequence of nonnegative constants with $\sum_{i=1}^{m} \alpha_i = 1$. Let $f(x) = \sum_{n=1}^{T} \alpha_n h_n(x)$, then margin is $\text{mar}_f(x,d) = f(x)d$. Then $| \text{mar}_f(x,d) | \leq 1$.
- Boosting works on examples that have small positive margin or negative margin. It increases the margin on the examples that have the smallest margin.
- $P(d \neq h(x)) \leq P(\text{mar}_f(x,d) \leq \theta) + O((d/(m\theta^2)^{1/2})$

Here generalization error bound is independent of $T$. 
References

Simple Practical Bayesian Learning

- Consider binary Bayesian classification problem where we attempt to learn from training data. Data is given by $X=(X_1, X_2, \ldots, X_n)$ and hypothesis is $D \in \{-1,1\}$
- MAP rule is $\text{argmax } P(D=i|X)$.
- Generative classifier: from training data learn prior, $P(D)$ and likelihood probability $P(X|D)$. Naïve Bayes.
- Discriminative classifier: from training data learn posterior, $P(D|X)$ directly. Logistic regression.
Naïve Bayes

- Prior is easy to learn from training data.
- Likelihood probability is more difficult to learn from training data as we need to learn likelihood distribution containing n+1 variables. Even if inputs are binary there are $2(2^n - 1)$ different possibilities to learn.
- To reduce computations assume that inputs variables are conditionally independent. Number of different possibilities is now $2n$.

$$\Pi_{j=1,n} P(X_j|D=i) = P(X_1, X_2, \ldots X_n|D=i)$$
Learning Prior and Likelihood Probabilities

- Assume inputs all take on a finite number of values then we can estimate prior and likelihood probabilities to make as estimate of the MAP decision rule
  \[ P(D=i \mid X_1, X_2, \ldots X_n) \propto P(D=i) \prod_{j=1,n} P(X_j\mid D=i) \]

- Priors: \( P(D=i) \approx \sum_{k=1,m} I(D_k =i)/m \)

- Likelihood:
  \[ P(X_j=l\mid D=i) \approx (\sum_{k=1,m} I(D_k =i, X_{kj}=l)+r)/(m+ 2r) \]
  \( r \) is added to smooth estimate. If \( r=0 \) and there are no counts, then likelihood is zero.
Naïve Bayes Comments

- Inputs can also be continuous. As an example inputs can be Gaussian where we work with mean and covariance.
- Studies have been conducted to test the effects of conditional independence assumption. We do not need to know posterior distribution, but only which hypothesis is more likely and in most cases conditional independence assumption works reasonably well.
- Naïve Bayes has been applied to many pattern recognition problems including text classification problems.
Logistic Regression

- Let \( P(D=1| X=x) = 1/(1+\exp(s)) \) where \( s=w^T x + w_0 \)
- Then negative log likelihood function given by
  \[-l(x) = \log \left( \frac{P(D=-1|X=x)}{P(D=1|X=x)} \right) = s \]
- Logistic regression is a linear classifier where if \( s \geq 0 \) decide \( D = -1 \) and otherwise decide \( D = 1 \).
- Common assumption is that likelihood functions are Gaussian vectors.
- Key is to learn weight vector based on training examples. Use log-likelihood energy function plus regularization and train using gradient descent.
Comparison between Naïve Bayes and Logistic Regression

- Naïve Bayes converges in $O(\log(n))$ updates, but asymptotic error more than Logistic Regression
- Logistic regression converges in $O(n)$ updates
- On tests of both algorithms common feature was for small number of examples Naïve Bayes better, but for large number of examples Logistic Regression better.