Chapter 7

Ensemble Classifiers

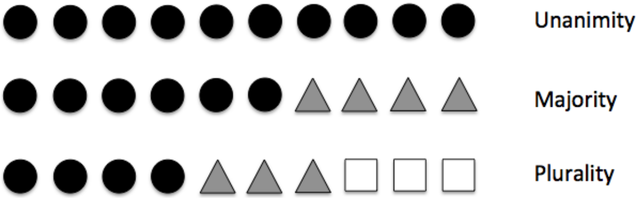
August 10, 2017

Learning with ensembles

- Our goal is to combined multiple classifiers
- Mixture of experts, e.g. 10 experts
- Predictions more accurate and robust
- Provide an intuition why this might work
- Simplest approach: majority voting

Majority voting

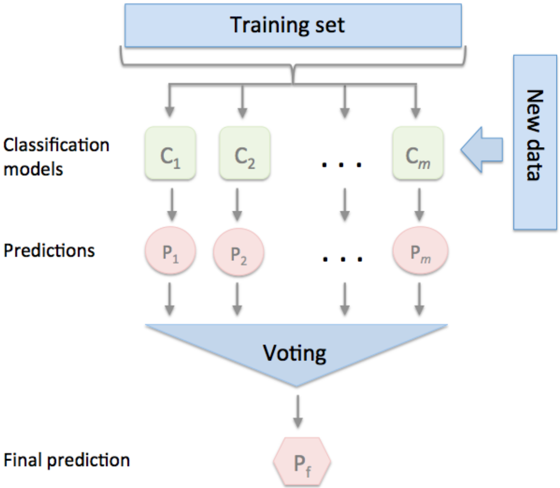
- Majority voting refers to binary setting
- Can easily generalize to multi-class: plurality voting
- Select class label that receives the most votes (mode)



Combining predictions: options

- Train m classifiers $C_1, \ldots, C_m$
- Build ensemble using different classification algorithms (e.g. SVM, logistic regression, etc.)
- Use the same algorithm but fit different subsets of the training set (e.g. random forest)

General approach



Combining predictions via majority voting

We have predictions of individual classifiers C_j and need to select the final class label $\hat{y}$

$$\hat{y} = mode\{C_1(\mathbf{x}), C_2(\mathbf{x}), \ldots, C_m(\mathbf{x})\}$$

For example, in a binary classification task where $class_1 = -1$ and $class_2 = +1$, we can write the majority vote prediction as follows:

$$C(\mathbf{x}) = sign\left[\sum_{j}^{m} C_j(\mathbf{x})\right] = \begin{cases} 1 & \text{if } \sum_j C_j(\mathbf{x}) \geq 0 \\ -1 & \text{otherwise} \end{cases}$$

Intuition why ensembles can work better

Assume that all n base classifiers have the same error rate ϵ. We can expresss the probability of an error of an ensemble can be expressed as a probability mass function of a binomial distribution:

$$P(y \geq k) = \sum_{k}^{n} \binom{n}{k} \epsilon^k (1-\epsilon)^{n-k} = \epsilon_{\text{ensemble}}$$

Here, $\binom{n}{k}$ is the binomial coefficient *n choose k*. In other words, we compute the probability that the prediction of the ensemble is wrong.

Example

Imagine we have 11 base classifiers ($n = 11$) with an error rate of 0.25 ($\epsilon = 0.25$):

$$P(y \geq k) = \sum_{k=6}^{11} \binom{11}{k} 0.25^k (1 - 0.25)^{11-k} = 0.034$$

So the error rate of the ensemble of $n = 11$ classifiers is much lower than the error rate of the individual classifiers.

Same reasoning applied to a wider range of error rates