CZ4041/CE4041: Machine Learning

Week 9: Ensemble Learning

Nerd Joke Time

- What does one support vector say to another support vector?
- I feel so marginalized.

Necessary Conditions

- Two necessary conditions for an ensemble classifier to perform better than a single classifier:
  1. The base classifiers are independent of each other
     - In practice, this condition can be relaxed that the base classifiers can be slightly correlated
  2. The base classifiers should do better than a classifier that performs random guessing (e.g., for binary classification, accuracy should be better than 0.5)

Question 1

Question 1

- Suppose there are 3 base classifiers, each classifier has error rate, $\varepsilon = 0.65$ or accuracy $\text{acc} = 0.35$
- Consider to combine the 3 base classifiers to make a prediction on a test instance using a majority vote
- Therefore, probability that the ensemble classifier makes a wrong prediction is:

$$\sum_{i=2}^{3} \binom{3}{i} \varepsilon^{i}(1-\varepsilon)^{3-i} = 3 \times 0.65^{2} \times 0.35 + 1 \times 0.65^{3} \times 1 = 0.71825$$

The Binomial Distribution

- This is the general expression for the binomial distribution.
- We perform N experiments. In each experiment, the probability of observing a negative outcome is ϵ. What is the probability for observing exactly M negative outcomes?

$$P(N, M) = \binom{N}{M} \varepsilon^{M} (1 - \varepsilon)^{N-M}$$

Applying That Formula …

- Suppose there are N (odd) independent base classifiers, each of which has the same error rate ε
  - Therefore, probability that the ensemble classifier makes a wrong prediction is:

There are i classifiers that make wrong predictions

$$P(N) = \sum_{i=\frac{N+1}{2}}^{N} \binom{N}{i} \varepsilon^{i} (1-\varepsilon)^{N-i}$$

The rest $N - i$ classifiers making correct prediction

At least more than half of the base classifiers

All possible combinations

Another Application

- Alternatively, let p be the probability that a single classifier makes the correct decision. The probability that the ensemble of $2n+1$ base classifier makes the correct decision is $P_c(2n+1)$

$$P_c(2n+1) = \sum_{m=n+1}^{2n+1} \binom{n}{\mathrm{m}} p^m (1-p)^{2n+1-m}$$

Question 1 (cont.)

- It can be proved that, with an odd number N
  - If $p > 0.5$ then $P_c(N)$ is monotonically increasing in N, and $P(N) \to 1$ as $N \to \infty$
  - If $p = 0.5$ then $P_c(N) = 0.5$ for all N
  - If $p < 0.5$ then $P_c(N)$ is monotonically decreasing in N, and $P(N) \to 0$ as $N \to \infty$
- Detailed proof can be found in the paper "Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance, 1997"

Detailed Proof

- This section is not required for the final exam.

Preparing for the Proof

- We only consider cases with odd number of base classifiers.
- We have two odd numbers, $2n - 1$ and $2n + 1$, where n is a non-negative integer.
- $P_C(m)$ is the probability that the ensemble of m base classifiers makes a correct decision.

Proof Sketch

- Very easy conceptually, though a bit tedious.
- Using the recurrence $\binom{n}{m} = \binom{n-1}{\mathrm{m}} + \binom{n-1}{m-1}$, we can relate $P_C(2n+1)$ to $P_C(2n)$
- Similarly, we can relate $P_C(2n)$ to $P_C(2n-1)$
- We want to show $P(2n+1) - P(2n-1) > 0$ if and only if $p > 0.5$

Preliminary

- $\binom{n}{n+1} = 0$ (you can't choose n+1 items from n items!)
- $\binom{n}{m} = \binom{n-1}{\text{m}} + \binom{n-1}{m-1}$ (recurrence from n items to n-1 items)
  - Intuitively, to choose *m* items from *n* items, we can single out one item *A*. If we decide not to choose *A*, then we need to choose *m* items from the remaining *n-1* items.
  - If we decide to choose *A*, then we will choose only *m-1* items from the remaining *n-1* items.
  - We add up the two possibilities.

Preliminary

- $\binom{n}{m} = \frac{n!}{(n-m)!\,m!}$
- $\binom{n-1}{m} + \binom{n-1}{m-1} = \frac{(n-1)!}{(n-1-m)!\,m!} + \frac{(n-m)!(m-1)!}{(n-1)!}$

$$= \frac{(n-1)!}{(n-1-m)!\,(m-1)!}\left(\frac{1}{m} + \frac{1}{n-m}\right)$$

$$= \frac{(n-1)!}{(n-1-m)!\,(m-1)!}\left(\frac{n}{m(n-m)}\right)$$

$$= \frac{n!}{(n-m)!\,m!} = \binom{n}{m}$$

First, we try to relate $P_C(2n+1)$ and $P_C(2n)$

$$
\begin{aligned}
&P_C(2n+1) \\
&= \sum_{m=n+1}^{2n+1} p^m (1-p)^{2n+1-m} \binom{2n+1}{m} \\
&= \sum_{m=n+1}^{2n+1} p^m (1-p)^{2n+1-m} \left[\binom{2n}{m} + \binom{2n}{m-1} \right] \\
&= (1-p) \sum_{m=n+1}^{2n+1} p^m (1-p)^{2n-m} \binom{2n}{m} \\
&\quad + p \sum_{m=n+1}^{2n+1} p^{m-1} (1-p)^{2n-(m-1)} \binom{2n}{m-1}
\end{aligned}
$$

We saw this just now

In this term, we have $\binom{2n}{2n+1} = 0$

$$P_C(2n+1) = (1-p)\sum_{m=n+1}^{2n} p^m (1-p)^{2n-m}\binom{2n}{m}$$ Notice the change of the upper limit

$$+ p\sum_{k=n}^{2n} p^k (1-p)^{2n-k}\binom{2n}{k}$$ Change of variable: $k = m - 1$

When $m = n + 1, k = n$

since $\binom{2n}{2n+1} = 0$

$$= (1-p+p)\sum_{k=n+1}^{2n} p^k (1-p)^{2n-k}\binom{2n}{k}$$

$$+ p^{n+1}(1-p)^n\binom{2n}{n}$$

$$= P_C(2n) + p^{n+1}(1-p)^n\binom{2n}{n},$$

$P_C(2n)$: An ensemble of $2n$ base classifiers makes the correct decision iff $n + 1$ base classifiers are correct

Next, we try to relate $P_C(2n)$ and $P_C(2n-1)$

$$
\begin{aligned}
&P_C(2n)\\
&= \sum_{m=n+1}^{2n} p^m(1-p)^{2n-m}\binom{2n}{m}\\
&= \sum_{m=n+1}^{2n} p^m(1-p)^{2n-m}\left[\binom{2n-1}{m}+\binom{2n-1}{m-1}\right]\\
&= (1-p)\sum_{m=n+1}^{2n} p^m(1-p)^{2n-m-1}\binom{2n-1}{m}\\
&\quad + p\sum_{m=n+1}^{2n} p^{m-1}(1-p)^{2n-m}\binom{2n-1}{m-1}
\end{aligned}
$$

Same trick
we saw earlier

Here we have $\binom{2n-1}{2n}$

$$P_C(2n) = (1-p)\sum_{m=n+1}^{2n-1} p^m (1-p)^{2n-1-m}\binom{2n-1}{m}$$

Change of the upper limit

$$+\, p\sum_{k=n}^{2n-1} p^k (1-p)^{2n-1-k}\binom{2n-1}{k}$$

Change of variable: $k = m - 1$

$$\text{since} \binom{2n-1}{2n} = 0$$

$$= (1-p+p)\sum_{k=n}^{2n-1} p^k (1-p)^{2n-1-k}\binom{2n-1}{k}$$

$$-\,(1-p)p^n(1-p)^{n-1}\binom{2n-1}{n}$$

$$= P_C(2n-1) - p^n(1-p)^n\binom{2n-1}{n}.$$

Putting things together, we can relate $P_C(2n+1)$ and $P_C(2n-1)$

$$P_C(2n+1) - P_C(2n-1) = p^n(1-p)^n \binom{2n-1}{n}(2p-1).$$

Proof: From Theorem 1,

$$\begin{aligned}
&P_C(2n+1) - P_C(2n-1) \\
&\quad = p^{n+1}(1-p)^n \binom{2n}{n} - p^n(1-p)^n \binom{2n-1}{n} \\
&\quad = p^n(1-p)^n \left\{ p\left[\binom{2n-1}{n} + \binom{2n-1}{n-1}\right] \right. \\
&\qquad \left. - \binom{2n-1}{n} \right\} \\
&\quad = p^n(1-p)^n \left[(p-1)\binom{2n-1}{n} + p\binom{2n-1}{n-1}\right] \\
&\quad = p^n(1-p)^n \binom{2n-1}{n}(p-1+p)
\end{aligned}$$

since $\binom{2n-1}{n-1} = \binom{2n-1}{n}$

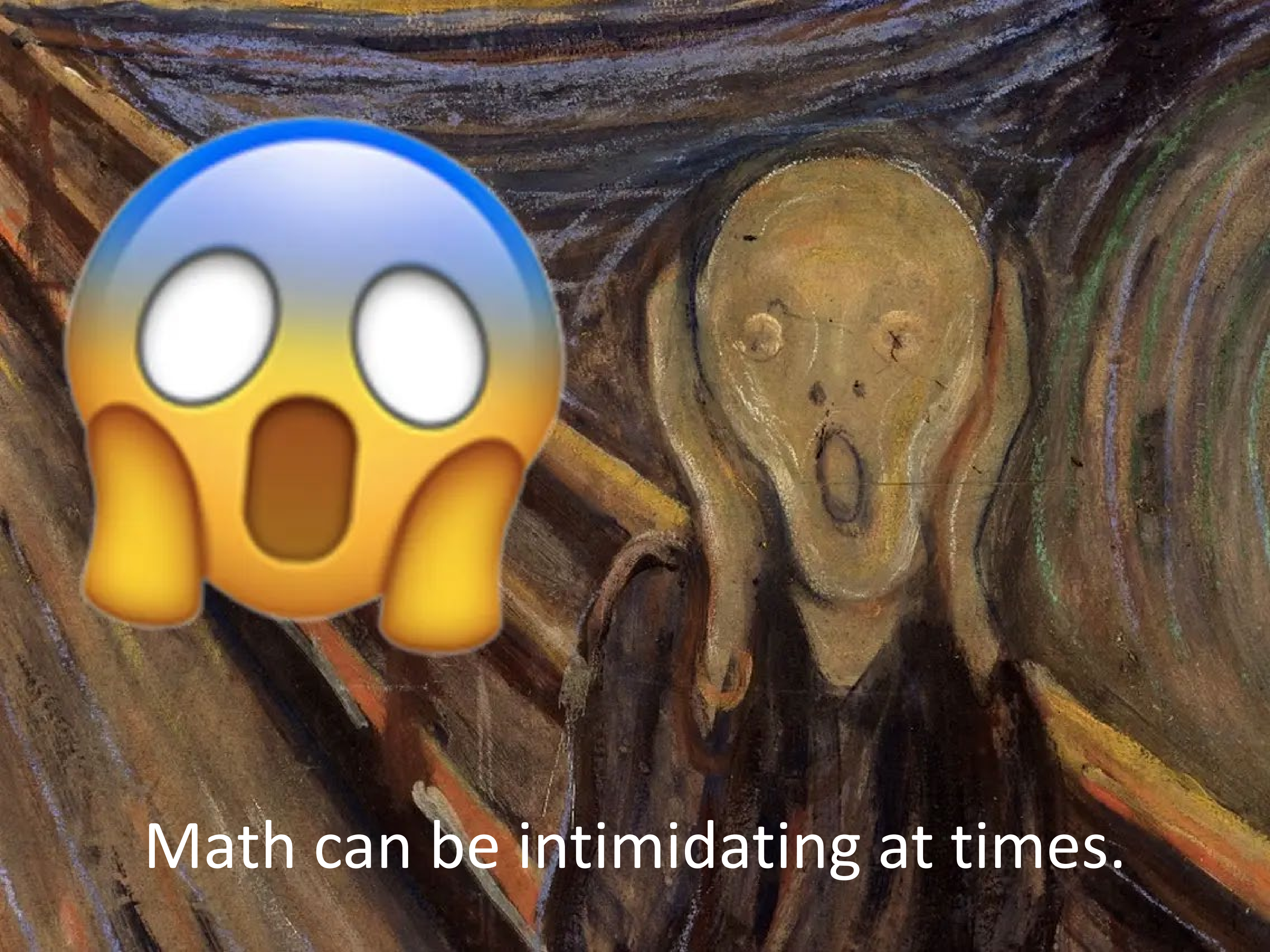
$$= p^n(1-p)^n \binom{2n-1}{n}(2p-1).$$

$$P_C(2n+1) - P_C(2n-1) = \underbrace{p^n(1-p)^n\binom{2n-1}{n}}_{\text{Always positive}}(2p-1).$$

When $p > 0.5, 2p - 1 > 0$

Therefore, $P_C(2n+1) > P_C(2n-1)$

The more base classifiers, the merrier.


Math can be intimidating at times.

KEEP CALM AND CARRY ON

保持冷静继续前进

சாந்தமாய் இரு. நிதானமாய் செயல்படு.

BERTENANG DAN TERUSKAN

Question 2

- Suppose we have trained 5 base binary classifiers: f1, f2, f3, f4 and f5. Their predictions on a validation dataset are shown in Table 1, where the last column denotes the ground-truth class labels. Which base classifiers would you choose to construct an ensemble learner?

ID	f_1	f_2	f_3	f_4	f_5	Ground Truth
P1	+	+	-	-	+	+
P2	+	+	-	+	-	+
P3	-	-	+	+	-	+
P4	-	-	+	-	+	+
P5	-	-	+	+	-	-
P6	-	-	-	+	+	+
P7	+	+	+	+	-	+
P8	-	+	+	-	+	-
P9	+	+	-	+	+	+
P10	-	-	-	+	-	-

Question 2

ID	f_1	f_2	f_3	f_4	f_5	Ground Truth
P1	+	+	-	-	+	+
P2	+	+	-	+	-	+
P3	-	-	+	+	-	+
P4	-	-	+	-	+	+
P5	-	-	+	+	-	-
P6	-	-	-	+	+	+
P7	+	+	+	+	-	+
P8	-	+	+	-	+	-
P9	+	+	-	+	+	+
P10	-	-	-	+	-	-
ACC:	0.7	0.6	0.4	0.60	0.60	
	✓	✓	✗	✓	✓	

- Almost perfectly correlated.
- On this dataset, f_1 strictly better than f_2

ID	f_1	f_2	f_3	f_4	f_5	Ground Truth
P1	+	+	-	-	+	+
P2	+	+	-	+	-	+
P3	-	-	+	+	-	+
P4	-	-	+	-	+	+
P5	-	-	+	+	-	-
P6	-	-	-	+	+	+
P7	+	+	+	+	-	+
P8	-	+	+	-	+	-
P9	+	+	-	+	+	+
P10	-	-	-	+	-	-
ACC:	0.7	0.6	0.4	0.60	0.60	