A Bit of Sequence Prediction: Lec 1

THE AI REVOLUTION

Next Token Prediction

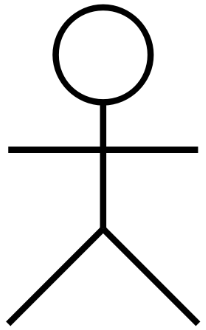
1. The workhorse of LLM's
2. Given tokens seen so far, predict the next one.
3. Studied under various names, autoregression, sequence prediction, etc.

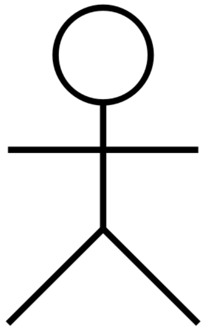
Next Token Prediction

1. The workhorse of LLM's
2. Given tokens seen so far, predict the next one.
3. Studied under various names, autoregression, sequence prediction, etc.
4. The simplest setting . . .

The Bit Prediction Game

Player I





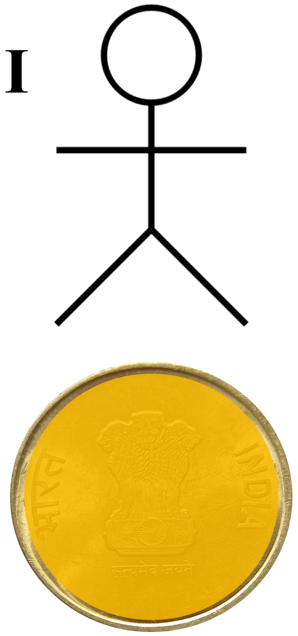
Player II

The Bit Prediction Game

1. Player I chooses heads or tails and does not reveal choice

The Bit Prediction Game

Player I



Player II

1. Player I chooses heads or tails and does not reveal choice
2. Player II chooses heads or tails

The Bit Prediction Game

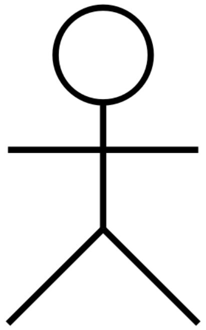
1. Player I chooses heads or tails and does not reveal choice
2. Player II chooses heads or tails
3. Coins are revealed, if coin faces match, player II gets both coins and if not player I gets both coinsx

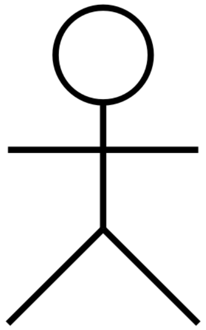
The Bit Prediction Game

1. Player I chooses heads or tails and does not reveal choice
2. Player II chooses heads or tails
3. Coins are revealed, if coin faces match, player II gets both coins and if not player I gets both coinsx

The Bit Prediction Game

Player I





Player II

1. Player I chooses heads or tails and does not reveal choice

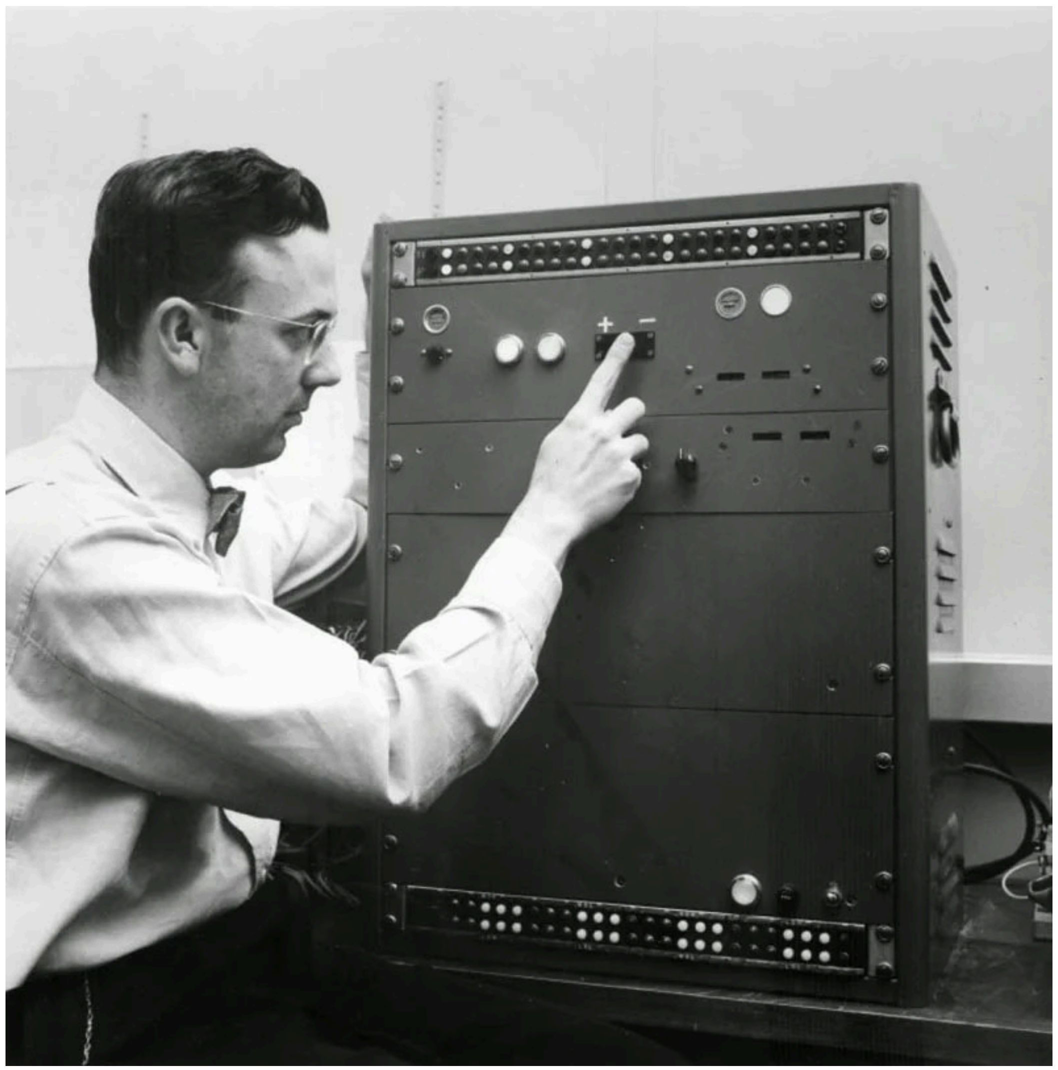
2. Player II chooses heads or tails

3. Coins are revealed, if coin faces match, player II gets both coins and if not player I gets both coinsx

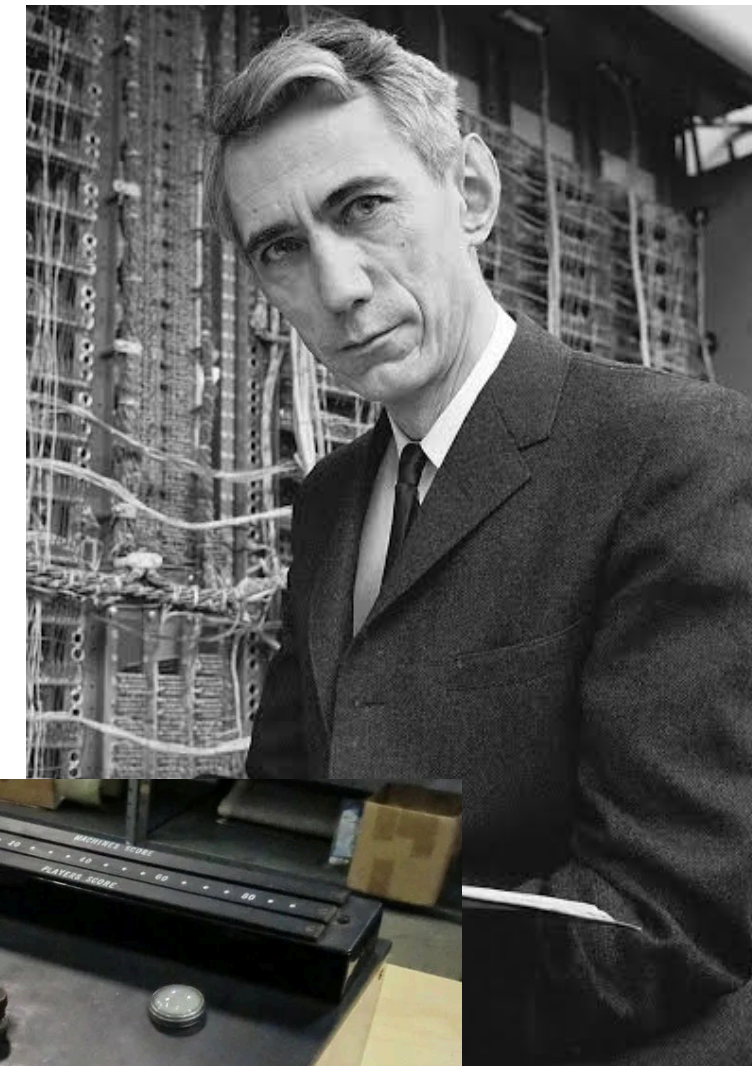
What is the optimal strategy?

Mind Reading Machine

David Hagelbarger



Claude Shannon



Made money playing with humans

Bit Prediction Framework

For $t = 1$ to n

Learner picks (possibly randomly) $\hat{y}_t \in \{\pm 1\}$

True outcome $y_t \in \{\pm 1\}$ is revealed

Learner suffers loss $\mathbf{1}\{\hat{y}_t \neq y_t\}$

End For

Bit Prediction Framework

For $t = 1$ to n

Learner picks (possibly randomly) $\hat{y}_t \in \{\pm 1\}$

True outcome $y_t \in \{\pm 1\}$ is revealed

Learner suffers loss $\mathbf{1}\{\hat{y}_t \neq y_t\}$

End For

Start simple: Goal, do well compared to majority in hindsight:

$$\mathrm{Reg}_n = \sum_{t=1}^{n} \mathbb{E}\left[\mathbf{1}\{\hat{y}_t \neq y_t\}\right] - \min_{b \in \{\pm 1\}} \sum_{t=1}^{n} \mathbf{1}\{b \neq y_t\}$$

Bit Prediction Framework

For $t = 1$ to n

Learner picks (possibly randomly) $\hat{y}_t \in \{\pm 1\}$

True outcome $y_t \in \{\pm 1\}$ is revealed

Learner suffers loss $\mathbf{1}\{\hat{y}_t \neq y_t\}$

End For

Start simple: Goal, do well compared to majority in hindsight:

$$\mathrm{Reg}_n = \sum_{t=1}^{n} \mathbb{E}\left[\mathbf{1}\{\hat{y}_t \neq y_t\}\right] - \min_{b \in \{\pm 1\}} \sum_{t=1}^{n} \mathbf{1}\{b \neq y_t\}$$

How well can we do w.r.t. this measure against minimax optimal strategy?

What is the strategy?

Strategies That Fail

1. Any deterministic strategy: Eg. majority so far. Why?
2. Randomized Prediction that predicts heads with probability equal to proportion of heads so far. Why?
3. Think sequence with first $n/3$ tails and the remaining heads.

So is $o(n)$ regret even possible?

Cover's Result

Lemma (T. Cover'65)

Let $\phi : \{\pm 1\}^n \mapsto \mathbb{R}$ *be any function s.t.,* $\forall i \in [n]$, *and* $y_1, \ldots, y_n$,

$$|\phi(y_1, \ldots, y_{i-1}, +1, y_{i+1}, \ldots, y_n) - \phi(y_1, \ldots, y_{i-1}, -1, y_{i+1}, \ldots, y_n)| \leq 1$$

then, there exists a randomized strategy such that for any sequence of bits,

$$\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t} \left[\mathbf{1}\{\hat{y}_t \neq y_t\} \right] \leq \phi(y_1, \ldots, y_n) \text{ if and only if } \mathbb{E}_{\epsilon} \phi(\epsilon_1, \ldots, \epsilon_n) \geq \frac{n}{2}$$

and further, the strategy achieving this bound on expected error is given by:

$$q_t = \frac{1}{2} + \frac{1}{2} \mathbb{E} \left[\phi(y_1, \ldots, y_{t-1}, -1, \epsilon_{t+1}, \ldots, \epsilon_n) - \phi(y_1, \ldots, y_{t-1}, +1, \epsilon_{t+1}, \ldots, \epsilon_n) \right]$$

where $\epsilon_1, \ldots, \epsilon_n$ *are Rademacher Random Variables.*

In plain Words

- For any sequence, expected number of mistakes made by forecaster $\leq \phi(\text{sequence})$ can be achieved.
- If and only such a result can be achieved against a random sequence.
- Caveat ϕ needs to satisfy stability condition that changing any one bit does not change its value by more than 1.

Cover's Result

Lemma (T. Cover'65)

Let $\phi : \{\pm 1\}^n \mapsto \mathbb{R}$ *be any function s.t.,* $\forall i \in [n]$*, and* $y_1, \ldots, y_n$*,*

$$|\phi(y_1, \ldots, y_{i-1}, +1, y_{i+1}, \ldots, y_n) - \phi(y_1, \ldots, y_{i-1}, -1, y_{i+1}, \ldots, y_n)| \leq 1$$

then, there exists a randomized strategy such that for any sequence of bits,

$$\sum_{t=1}^{n} \mathbb{E}_{\hat{y}_t \sim q_t} \left[\mathbf{1}\{\hat{y}_t \neq y_t\} \right] \leq \phi(y_1, \ldots, y_n) \text{ if and only if } \mathbb{E}_{\epsilon} \phi(\epsilon_1, \ldots, \epsilon_n) \geq \frac{n}{2}$$

and further, the strategy achieving this bound on expected error is given by:

$$q_t = \frac{1}{2} + \frac{1}{2} \mathbb{E}\left[\phi(y_1, \ldots, y_{t-1}, -1, \epsilon_{t+1}, \ldots, \epsilon_n) - \phi(y_1, \ldots, y_{t-1}, +1, \epsilon_{t+1}, \ldots, \epsilon_n) \right]$$

where $\epsilon_1, \ldots, \epsilon_n$ *are Rademacher Random Variables.*

Cover's Result Proof

- The if direction is trivial.
- Only if direction: Plug in q_t recursively starting from n
- Idea for deriving q_t: Solve minimax optimization starting from n backwards
- Why was the condition on ϕ needed?

Bit Prediction

- Back to goal to minimize:

$$\mathrm{Reg}_n = \sum_{t=1}^{n} \mathbb{E}\left[\mathbf{1}\{\hat{y}_t \neq y_t\}\right] - \min_{b\in\{\pm 1\}} \sum_{t=1}^{n} \mathbf{1}\{b \neq y_t\}$$

- Pick $\phi(y_1, \ldots, y_n) = \min_{b\in\{\pm1\}} \sum_{t=1}^{n} \mathbf{1}\{b \neq y_t\} + C_n$

- Condition $\mathbb{E}\left[\phi(\epsilon_1, \ldots, \epsilon_n)\right] = \frac{n}{2}$ yields:

$$\begin{aligned} C_n &= \frac{n}{2} - \mathbb{E}\left[\min_{b\in\{\pm1\}} \sum_{t=1}^{n} \mathbf{1}\{b \neq \epsilon_t\}\right] \\ &= \frac{n}{2} - \frac{1}{2}\mathbb{E}\left[\min_{b\in\{\pm1\}} \sum_{t=1}^{n} (1 - b\cdot\epsilon_t)\right] \\ &= \frac{1}{2}\mathbb{E}\left[\max_{b\in\{\pm1\}} b\cdot\left(\sum_{t=1}^{n}\epsilon_t\right)\right] = \frac{1}{2}\mathbb{E}\left[\left|\sum_{t=1}^{n}\epsilon_t\right|\right] \le \frac{\sqrt{n}}{2} \end{aligned}$$

Bit Prediction: Rademacher Complexity

- Let $\mathcal{F} \subset \{\pm 1\}^n$ be a set of benchmarks we want to compete with.
- Consider the goal of minimizing regret:

$$\mathrm{Reg}_n(\mathcal{F}) = \sum_{t=1}^{n} \mathbb{E}\left[\mathbf{1}\{\hat{y}_t \neq y_t\}\right] - \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \mathbf{1}\{f_t \neq y_t\} \quad ,$$

- One can use $\phi(y_1, \ldots, y_n) = \min_{f \in \mathcal{F}} \sum_{t=1}^{n} \mathbf{1}\{f_t \neq y_t\} + C_n(\mathcal{F})$
- First, ϕ satisfies stability condition (easy to verify)!
- Using same steps as previous case we get:

$$C_n(\mathcal{F}) = \frac{1}{2}\mathbb{E}\left[\max_{f \in \mathcal{F}} \sum_{t=1}^{n} f_t \epsilon_t\right] \leq O(\sqrt{n \log |\mathcal{F}|})$$

Bit Prediction: Rademacher Complexity

- The term $\mathbb{E}\left[\max_{f\in\mathcal{F}} \sum_{t=1}^{n} f_t \epsilon_t\right]$ is referred to as Rademacher Complexity in Statistical Learning Theory (SLT)

- In SLT input instances x's and y's are drawn iid from fixed distribution and goal is excess risk.

- If we had a priroi $x_1, \ldots, x_n$ and then allowing adversary to pick labels y's as we go, then one can still use Cover's result using $\mathcal{F} = \{f(x_1), \ldots, f(x_n) : f \in \mathcal{F}\}$

- In fact if one has access to unlabeled data/context drawn iid from fixed distribution, one can still use this result with strategy:

$$q_t = \frac{1}{2} + \frac{1}{2}\mathbb{E}\Big[\phi(x_1, y_1, \ldots, x_{t-1}y_{t-1}, x'_t, -1, x'_{t+1}, \epsilon_{t+1}, \ldots, x'_n\epsilon_n) - \phi(x_1, y_1, \ldots, x_{t-1}y_{t-1}, x'_t, +1, x'_{t+1}, \epsilon_{t+1}, \ldots, x'_n\epsilon_n)\Big]$$

What if we had context that depended on our past predictions?

Think bit prediction where contexts are past y's and we want to compete with strategies that take these strategies into account.

Does Cover's result work as is, if not what cracks?