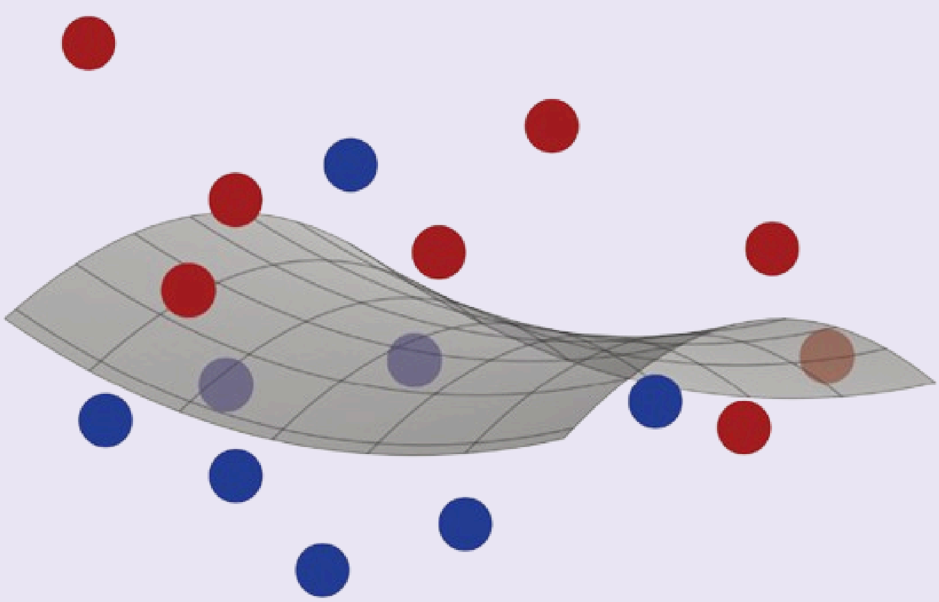
Foundations of Machine Learning

DAY - 10

Guarantees for Finite Hypothesis Sets – Inconsistent Case



Guarantees for Finite Hypothesis Sets – Inconsistent Case

Created By: **Birva Dave**
medium.com/@birva1809
github.com/Birva1809
linkedin.com/in/birva-dave
birvadave1809@gmail.com

- In real-world scenarios, it's common that no hypothesis in the set H fits all training examples perfectly — especially when:
  - The learning problem is hard.
  - The hypothesis set is too simple to capture the true concept.
- Still, inconsistent hypotheses with low training error can generalize well, and we can prove it using probabilistic guarantees, specifically Hoeffding's inequality.

Hoeffding's Inequality — Key Corollary

- For any fixed hypothesis h: X → {0, 1}, and any ϵ>0, the following inequalities hold:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\widehat{R}_S(h) - R(h) \geq \epsilon\right] \leq \exp(-2m\epsilon^2)$$

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\widehat{R}_S(h) - R(h) \leq -\epsilon\right] \leq \exp(-2m\epsilon^2)$$

- By the union bound, this implies the following two-sided inequality:

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\left|\widehat{R}_S(h) - R(h)\right| \geq \epsilon\right] \leq 2\exp(-2m\epsilon^2)$$

- This means: the more training samples m, the closer the empirical error $R_S(h)$ is to the true error R(h), with high probability.

Guarantees for Finite Hypothesis Sets – Inconsistent Case

Created By: **Birva Dave**
medium.com/@birva1809
github.com/Birva1809
linkedin.com/in/birva-dave
birvadave1809@gmail.com

Generalization Bound for a Single Hypothesis

- Let h be any fixed hypothesis from H. Then, for any δ>0, with probability at least 1−δ:

$$R(h) \leq R_S(h) + \sqrt{\frac{1}{2m} \log \frac{2}{\delta}}$$

- This tells us that the true error is not much more than the training error, especially if m is large.

Example: Tossing a Biased Coin

- Suppose the true probability of heads is ppp, and we always guess tails.
- Then:
  - True error = p
  - Empirical error = observed fraction of heads in the sample
- Hoeffding's inequality ensures that the observed proportion p^ is close to p with high probability.
- Example bound:
- If δ=0.02, m=500, then with 98% confidence:

$$|p - \hat{p}| \leq \sqrt{\frac{\log(2/\delta)}{2m}} \approx 0.048$$

Guarantees for Finite Hypothesis Sets – Inconsistent Case

Created By: **Birva Dave**
medium.com/@birva1809
github.com/Birva1809
linkedin.com/in/birva-dave
birvadave1809@gmail.com

Can We Apply This to Learned Hypotheses Like hSh_ShS?

- Not directly — because h_S is chosen based on the sample S, so it's not fixed.
- Instead, we must use a uniform convergence bound that holds for all hypotheses in H simultaneously.

Generalization Bound for All Hypotheses in Finite H (Inconsistent Case)

- Let H be a finite hypothesis class. Then with probability at least 1−δ, for every h∈H:

$$R(h) \leq R_S(h) + \sqrt{\frac{1}{2m}\left(\log|H| + \log\frac{2}{\delta}\right)}$$

Interpretation

- The bound applies uniformly over all hypotheses — even if h_S was picked based on S.
- The bound is slightly looser than in the consistent case — the second term grows with the square root of log ∣ H, instead of linearly.
- Therefore, to achieve the same guarantee as the consistent case, you'd need a quadratically larger sample.

Guarantees for Finite Hypothesis Sets – Inconsistent Case

Created By: **Birva Dave**
medium.com/@birva1809
github.com/Birva1809
linkedin.com/in/birva-dave
birvadave1809@gmail.com

Trade-Off: Empirical Error vs. Hypothesis Set Size

- A larger hypothesis class can reduce training error but increases the risk of overfitting.
- A smaller class may generalize better if it still fits the data well.
- This embodies Occam's Razor: prefer simpler explanations (hypothesis sets) when all else is equal.

Summary: Inconsistent Case vs. Consistent Case

Property	Consistent Case	Inconsistent Case
Requires zero training error	Yes	No
Bound form	$R(h_S) \leq \frac{1}{m}\left(\log\lvert H\rvert + \log\frac{1}{\delta}\right)$	$R(h) \leq R_S(h) + \sqrt{\frac{1}{2m}\left(\log\lvert H\rvert + \log\frac{2}{\delta}\right)}$
Data requirement	Lower	Higher
Dependence on	log ∣ H∣, log(1/δ), 1/m	RS(h), log(1/δ), 1/m, √(1/2m)
Learning feasible?	Yes, if consistent hypothesis exists	Yes, if empirical error is low and sample is large enough