1.Overview

1. Asymptotics

Setting: Given data drawn based on unknown parameter θ^*, we compute the estimate $\hat{\theta}$ from data. How close is $\hat{\theta}$ to θ^*?

(i) For Gaussian models and fixed design linear regression, we can compute $\hat{\theta} - \theta^*$ in closed form.

(ii) For most models, we can't compute $\hat{\theta} - \theta^*$ directly. But we can use asymptotics, whose idea is to take Taylor expensions and show asymptotic normality: $\sqrt{n}(\hat{\theta} - \theta^*) \rightarrow \mathcal{N}(\mu, \sigma^2)$ $(n \to \infty)$.

(iii) Maximum likelihood estimators play a significant role in our analysis. An old approach is brought to bear on the local optima problem.

---

2. Uniform convergence

Drawbacks of asymptotics:

1. Smoothness assumption: Invalid when analyze the hinge loss.
2. we don't know how large n has to be.

Setting (Uniform converge):

Training set: (x, y) pairs, learning algorithm chooses a predictor $h: \mathcal{X} \rightarrow \mathcal{Y}$ fom a hypothesis class $\mathcal{H}$. We evaluate it based on test data. Q: How do training error $\hat{L}(h)$ and test error $L(h)$ relater to each other?

(i) For a fixed $h \in \mathcal{H}$, $\hat{L}(h)$ is an average of i.i.d. los

(i) For a fixed $h \in \mathcal{H}$, $L(h)$ is an average of i.i.d. r.v.,
by Hoeffding's ineq., $\hat{L}(h) \to L(h)$.

(ii) Consider the empirical risk minimizer (ERM):

$$\hat{h}_{ERM} \in \arg\min_{h \in \mathcal{H}} \hat{L}(h)$$

Can we argue the relationship between $\hat{L}(\hat{h}_{ERM})$ and $L(\hat{L}_{ERM})$?

The key is: $\hat{h}_{ERM}$ depends on $\hat{L}$ (i.e., the training data)

We will show (using uniform convergence):

$$L(\hat{h}_{ERM}) \leq \hat{L}(\hat{h}_{ERM}) + O_p\left(\sqrt{\frac{\text{Complexity}(\mathcal{H})}{n}}\right)$$

(iii) We will get distribution-free results.

---

3. Kernel methods

To think what models should be learned?

Setting:

A regression task: predicting $y \in \mathbb{R}$ from $x \in \mathcal{X}$. We define a positive semidefinite kernel $k(x, x')$, which capture the 'similarity' between x and x', then define $f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x$

Finally, we difine the reproducing kernel Hilbert space (RKHS).

---

4. Online learning

The world is a dynamic place,

(i) data points might be dependent (not i.i.d)

(ii) data might be arriving in a stream (not in a batch)

Setting:

The online learning setting is a game between a learn and nature:

Iteration $t = 1, \cdots, T$

* Learner recieves input x_t
* Learner outputs prediction p_t
* Learner recieves true label y_t
* (Update)

How do we evaluate?

- Loss function
- Let $\mathcal{H}$ be a set of fixed expert predictors
- Regret : we will show $\text{Regret} \le O\sqrt{T \log |\mathcal{H}|}$

Online learning always leads to MAB setting.