Lecture 5. Regularization & Model Selection.

- Introduction:
  (1) different model → learning problem.
  (2) finite sets of models.
  e.g. linear model, SVM, CNN.

- Cross validation:
  (1) empirical risk minimization.
  (2) simple cross validations:
  ① training set S:
  split into S_{train} / S_{cv}
  ② train on S_{train}
  ③ select M_i:
  $\min \hat{\varepsilon}_{S_{cv}}(h_i)$ (retrain M_i)
  *: sensitive model: no retrain.
  (3) drawback: wasting data (30%)
  data is scarce (e.g. $m=30$)
  promoted idea: k-fold.
  ① S: randomly split into disjoint subset
  $S_1, S_2, S_3, \cdots S_k$.
  ② For $j=1,2,\cdots k$.

train M_i on $S / S_j \Rightarrow h_{ij}$

test on $S_j \Rightarrow \hat{\varepsilon}_{S_j}(h_{ij})$

③ pick $\min_{M_i} \sum_{j=1}^{k} \hat{\varepsilon}_{S_j}(h_{ij})$

*: retrain on the whole set.

(computational expensive)

(4) leave-one-out:

only hold out one training example.

◦ Feature Selection:

(1) small number of features → relevant.

n features → 2^n subsets. → heuristic search.

(2) forward search.

① $F = \Phi$ (feature set)

② Repeat:

$i = 1, 2, \cdots n$ and $i \notin F$

find $F_i = F \cup \{i\}$ optimal.

③ Output F. ($O(n^2)$)

Similarly: backward.

(3) Filter feature selection:

score $S(i)$ → how informative.

mutual information:

$$MI(X_i, y) = \sum_{x_i} \sum_{y_i} P(X_i, y) \log \frac{P(X_i, y)}{P(X_i) P(y)}$$

$KL(P(X_i, y) \| P(X_i)P(y))$

(how different distributions are)

rank → choose k (using CV)

- Bayesian statistics & regularization:

(1) maximum likelihood:

$$\theta_{ML} = \arg\max_{\theta} \prod_{i=1}^{m} P(y_i | X_i, \theta)$$

view θ as an unknown parameter.

constant value but unknown.

θ not random.

task: statistical procedure → frequentist.

→ estimate parameter θ.

(2) Bayesian View:

θ is random

task: prior belief. → posterior

$$P(\theta | S) = \frac{P(S|\theta)P(\theta)}{P(S)}$$

$$= \frac{\left(\prod_{i=1}^{m} P(y_i | X_i, \theta)\right) P(\theta)}{P(S)}$$

usually: $\theta \sim N(0, \tau^2 I)$

(3) Maximum a Posteriori.

$$\theta_{MAP} = \arg\max_{\theta} \prod_{i=1}^{m} P(y_i | X_i, \theta) P(\theta)$$

smaller norm ⇒ less likely to overfit.