Cost Sensitive Learning

CS 550: Machine Learning

Cost Sensitive Learning

- Many different types of cost may occur in real-world applications
  1. Cost of misclassification error
  2. Cost of feature extraction/representation
  3. Cost of computation
  4. Cost of teacher (also related with active learning)
  5. Other costs (see Turney's paper)

P. Turney, "Types of cost in inductive concept learning," Workshop on Cost-Sensitive Learning at the 17th International Conference on Machine Learning, (ICML), 2000.

Cost of Misclassification Error

- Different costs may be associated with different types of misclassification

$$R(\alpha_i \mid x) = \sum_{j=1}^{C} P(C_j \mid x) \cdot \underbrace{\lambda(\alpha_i \mid C_j)}$$

You may use the loss function to reflect this cost. Then select the class, for which the conditional risk is minimum.

Cost of Feature Extraction

- Select a subset of available features at the beginning
  - Feature subset is the same for all samples
  - How can you incorporate the cost of feature extraction into feature selection?

- Dynamically select the next feature
  - Different feature subsets might be used for different samples
  - Have to decide where to stop
  - Suppose that you have already extracted a set of features F_{prev} and want to decide whether or not you should extract another feature F. How can you decide it? And if so, how can you select the next feature?

Example: Dynamically select the feature set

- Define an expected utility EU of using F_{prev} for sample x

$$EU(F_{prev} \mid x) = \sum_{j=1}^{C} \underbrace{P(C_j \mid F_{prev}, x)}_{\text{Posterior of obtaining class } C_j \text{ when } F_{prev} \text{ is used for sample } x} \underbrace{u(C_j, F_{prev})}_{\text{Utility of using this feature set and obtaining class Cj}} - \beta \, \text{cost}(F_{prev})$$

e.g., 1 for correct classification and 0 for misclassifiaction

Here we use a linear function to combine the utility and the feature extraction cost. You may also combine them differently.

- Now define the EU of using an additional feature F

$$EU(F_{prev} + F \mid x) = \sum_{j=1}^{C} P(C_j \mid F_{prev} + F, x) \; u(C_j, F_{prev} + F) - \beta \, \text{cost}(F_{prev}) - \beta \, \text{cost}(F)$$

- Use the feature that maximizes the net EU
- Stop if the net EU is negative for all features

Example: Dynamically select the feature set

- Net expected utility of using feature F

$$net(F \mid x) = EU(F_{prev} + F \mid x) \; - \; EU(F_{prev}\, x)$$

$$= \left[\sum_{j=1}^{C} P(C_j \mid F_{prev} + F\,,\, x)\; u(C_j\,,\, F_{prev} + F) \; - \; \beta\, \text{cost}(F_{prev}) \; - \; \beta\, \text{cost}(F) \right] \; -$$

$$\left[\sum_{j=1}^{C} P(C_j \mid F_{prev}\,,\, x)\; u(C_j\,,\, F_{prev}) \; - \; \beta\, \text{cost}(F_{prev}) \right]$$

$$= \sum_{j=1}^{C} P(C_j \mid F_{prev} + F\,,\, x)\; u(C_j\,,\, F_{prev} + F) \; - \; \sum_{j=1}^{C} P(C_j \mid F_{prev}\,,\, x)\; u(C_j\,,\, F_{prev}) \; - \; \beta\, \text{cost}(F)$$

How do you estimate $P(C_j \mid F_{prev} + F, x)$ without extracting this feature F?

Cost of Computation

- Static complexity
  - E.g., the code size
- Dynamic complexity
  - Time complexity (training or testing time complexity)
  - Space complexity (e.g., memory required for a kNN classifier)
- Decide which tests/classifiers are to be used considering the dynamic complexity
  - Bayesian decision theory could also be used to formulate this problem (similar to our previous feature selection example)
  - Select them at the beginning for all samples or dynamically select them for each sample

Cost Sensitive Learning

- How to combine different types of cost?
- How to select features when you have a budget (limit for each sample or average budget limit)?
- How to select classifiers when you have a maximum waiting time?
- How to minimize the cost of a teacher? (active learning – next topic)