T-61.3050 Machine Learning: Basic Principles

Exercise 2 - Supervised Learning

Department of Information and Computer Science
Aalto University School of Science

September 19, 2012

Problem 1 – Dummy model

- We are given data $(x_i, r_i), i = 1, \ldots, N$ to train the model
- The model returns predictions $\hat{r}$
- Dummy model purpose – baseline model
- Predictions $\hat{r}_j$ are independent of inputs x_j
- In some cases, very basic and simplistic predictions are more than enough for the problem at hand
- Simplest case: always guess 0 or always guess 1

Problem 1 – Classification

- Let N_0 and N_1 be the number of times that $r_i = 0$ and $r_i = 1$ respectively, on the training data
- We use the error function $E(r, \hat{r}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}(r_i \neq \hat{r}_i)$
- If we always answer 0 the error is $E(r, \mathbf{0}) = \frac{1}{N} N_1 = \frac{N_1}{N}$
- If we always answer 1 then $E(r, \mathbf{1}) = \frac{N_1}{N}$

Problem 1 – Classification simple example

- If the class distribution is *skewed*, so that $N_0 >> N_1$, then the dummy model answering 0 all the time will be correct very often.
- For example: class 0: male, class 1: female
- At certain campuses $N_0/N_1 = 100$
- Dummy model error: $\frac{N_1}{N_0+N_1} = \frac{N_0/100}{N_0+N_0/100} \approx 0.0099$
- If the features used by the non-dummy model are not very predictive, e.g. gender classification from the color of clothing, then beating the dummy model can be hard
- With a skewed label distribution there are few examples for the minor class, and therefore learning its properties are difficult

Problem 1 – Regression

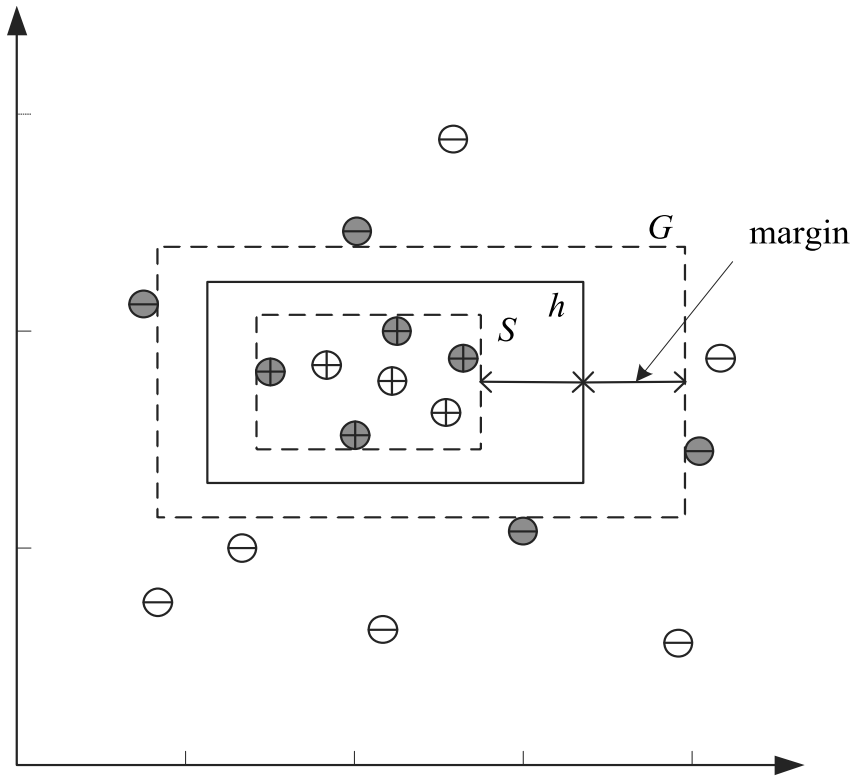
- Now our error function is $E(r, \hat{r}) = \sum_{i=1}^{N} (r_i - \hat{r}_i)^2$
- Our dummy model just returns a single value $\hat{r}_i = m$
- What should m be?

Problem 1 – Regression

- If we minimize $E(r, \hat{r}) = \sum_{i=1}^{N}(r_i - m)^2$ w.r.t. m:
- $\frac{dE}{dm} = \sum_{i=1}^{N} -2(r_i - m) = -2\sum_{i=1}^{N}(r_i - m) = 2Nm - 2\sum_{i=1}^{N} r_i$
- Find zeros:
  $\frac{dE}{dm} = 0 <=> 2Nm = 2\sum_{i=1}^{N} r_i <=> m = \frac{1}{N}\sum_{i=1}^{N} r_i$
- The best m is the mean of r

Problem 2 – Hypotheses

- Most specific hypothesis (S) and most general hypothesis (G)
- Depending on the problem, we will pick some hypothesis $h \in \mathcal{H}$ in version space



Problem 2 – Hypotheses

Prior information available:

1. Reasons for choosing either S or G can be goal-driven:
   - Reducing false positives $\rightarrow S$ (example virus detection)
   - Reducing false negatives $\rightarrow G$ (example disease detection)
2. Class distortion: one class dominates the other

Problem 2 – Hypotheses

- Without prior information – average over S and G
- Maximizing the margin between classes
- Similar inputs have similar outputs:
  - input X_j closer to $S \rightarrow f(X_j) = 1$
  - input X_j closer to $G \rightarrow f(X_j) = 0$
- Training and test samples are assumed to be drawn from same distribution $\rightarrow$ perturbed training samples should have same label
- Theoretical justification for maximizing class separation (margin)

Problem 2 – Hypotheses

- Intuition for taking average of S and G: resampling
  1. remove samples on the boundary of S or G
  2. update boundaries
  3. how will the samples that were removed from training be classified?
- To minimize false negatives and positives, average is less sensitive to resampling

Problem 3 – Linear Regression

Blackboard stuff

Problem 3 – Linear Regression

$$E(w_1, w_0|\mathcal{X}) = \frac{1}{N}\sum_{i=1}^{N}\left[r^t - (w_1 x^t + w_0)\right]^2$$

$$\begin{aligned}
\frac{\partial E}{\partial w_0} &= \sum_t \left[r^t - (w_1 x^t + w_0)\right] = 0 \\
\sum_t r^t &= w_1 \sum_t x^t + N w_0 \\
w_0 &= \sum_t r^t/N - w_1 \sum_t x^t/N = \overline{r} - w_1 \overline{x}
\end{aligned}$$

Problem 3 – Linear Regression

$$
\begin{aligned}
\frac{\partial E}{\partial w_1} &= \sum_t \left[r^t - (w_1 x^t + w_0) \right] x^t = 0 \\
\sum_t r^t x^t &= w_1 \sum_t (x^t)^2 + w_0 \sum_t x^t \\
\sum_t r^t x^t &= w_1 \sum_t (x^t)^2 + (\overline{r} - w_1 \overline{x}) \sum_t x^t \\
\sum_t r^t x^t &= w_1 \left(\sum_t (x^t)^2 - \overline{x} \sum_t x^t \right) + \overline{r} \sum_t x^t \\
\sum_t r^t x^t &= w_1 \left(\sum_t (x^t)^2 - \overline{x} N \overline{x} \right) + \overline{r} N \overline{x} \\
w_1 &= \frac{\sum_t r^t x^t - \overline{x} \overline{r} N}{\sum_t (x^t)^2 - N \overline{x}^2}
\end{aligned}
$$

Tips for home assignment 1:

- Same as demonstration 3 but with vectors. We take W, R, and Φ as in slide 32 of the second lecture:
  1. R is a $1 \times N$ vector, containing the targets $(r^1 r^2 \cdots r^N)$
  2. W is a $1 \times k$ vector, containing the weights to be optimized.
  3. Φ is an $k \times N$ matrix with elements $\Phi_{i,t} = \phi_i(x^t)$ (i.e. the t^{th}column vector $\phi_i(x^t) = (1, x_1^t, x_2^t, \ldots, x_{k-1}^t)^T$)

  For obtaining the least squares solution for this multilinear regression problem we should optimize $w_1, \ldots, w_k$ such that the error

$$E(g|\Phi) = \frac{1}{N}\sum_{t=1}^{N}(r_t - g(\Phi))^2 = \frac{1}{N}\sum_{t=1}^{N}(r_t - W\Phi_{\cdot t}))^2$$

  is minimized. In matrix notation this becomes

$$\frac{1}{N}(R - W\Phi)(R - W\Phi)^T$$

- Deriving matrices: see The Matrix Cookbook
- Don't be confused by basis functions; they are constant wrt W and thus don't affect the derivation

Tips for home assignment 2:

- use

  `load()`

  to load in the data
- selecting samples for training/validation:

  ```
  perm = randperm(n_data);
  ```
- and to select a subset of indexes from vector perm, useful to know that you can do

  ```
  training_idx = perm([1:i-1 i+1:end]);
  test_idx = perm(i);
  ```

  which also works with the first and last elements

Tips for home assignment 3:

- What is a scatter plot? See

  `http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient`

  for a good example
- useful plotting commands:

  `plot(a,b,'.'), hist(), mean()`