Pattern Analysis. Lab 3. Bonus.

Konstantin Tretyakov

November 9, 2006

Risk Estimation

Consider the problem of supervised learning, regression in particular. We are given a dataset and we shall use it to find a function f that will perform "nicely" in the future. What does it mean — to perform "nicely"? It means the function should have, on average, a small error. Note that we are not only speaking about the error we observe on the *training set* given to us, and not even the error on the *test set*, but rather the error "in general", which is a completely different thing. Why should a small error on the training or test set necessarily result in a small error overall? Can we estimate this overall error somehow?

Consider a one-dimensional regression task $Y = f(X)$, where $X \sim \mathcal{U}(0,1)$ is distributed uniformly and for each X the distribution of Y is gaussian with mean $\mu = X(2-X)$ and standard deviation $\sigma = 0.2$: $Y \sim \mathcal{N}(X(2-X), \sigma)$. Let's generate a sample $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ from this distribution, fit a *linear* regression model to it and analyze the *overall average error (risk)* of this model.

Let f be some regression function. We define its *risk* as follows:

$$R(f) = \int (f(x) - y)^2 dF(x, y),$$

that is, it's the *mean squared error* of the function, where the mean is taken over *all possible pairs* (x, y). In a sense, it corresponds to the average squared error of our function on an infinite dataset.

A "good" function will minimize this risk $R(f)$. Which one is good? In this simple example it's possible to see that the best regression function is $f(x) = x(2-x)$, and its risk is $\sigma^2 = 0.04$.

However, we'll be fitting our data with a *linear* function $f(x) = ax + b$ instead and certainly it's risk will be worse. It is possible to show, that a risk of a linear f will be equal to:

$$R(f) = \sigma^2 + \frac{1}{5} + \frac{a-2}{2} + \frac{2b + (a-2)^2}{3} + (a-2)b + b^2 \qquad (1)$$

Exercise 1*: Prove it.
Hint: show that

$$R(f|x_0) = \int_y (f(x) - y)^2 dF(y|x_0) = \sigma^2 + (f(x_0) - x_0(2 - x_0))^2$$

and use the equation $R(f) = \int R(f|x)dx$.

In real life we can never know the true risk exactly, so we have to *estimate* it from data. The empirical estimate of risk is, of course, the mean squared error[1]:

$$\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} l_i, \qquad \text{where} \quad l_i = (f(x_i) - y_i)^2$$

This gives us a so-called *point estimate*. To characterize the precision of this estimate we can construct a *confidence interval*. We use the fact, that for a reasonably large n the variable $\hat{R}(f)$ distributed normally with mean $R(f)$ and standard deviation $\frac{\sigma_R}{\sqrt{n}}$ where σ_R is the standard deviation of $(f(x)-y)^2$. Therefore a 95% confidence interval for $R(f)$ is

$$\hat{R}(f) \pm 1.96\frac{\hat{\sigma}_R}{\sqrt{n}}, \qquad \text{where} \quad \hat{\sigma}_R = \frac{1}{n-1}\sum_i (l_i - \hat{R}(f))^2$$

Exercise 2: Generate a training set of 500 points. Estimate a linear regression model. Plot the points and the regression line. Compute the risk of the model by using the equation (1). Generate a test set of 500 points. Use it to estimate risk empirically. Find a 95% confidence interval. Does the true risk fall into the interval? Repeat the experiment several times. How often the true risk is not in the interval? How much do we underestimate risk if we use the training set to estimate it instead of the test set?
Hints:

- Generating data:

```
rand("uniform");
X = rand(n, 1);
rand("normal");
y = X.*(2 - X) + s * rand(X);
```

- Linear regression

```
w = [ones(n, 1) X]\y;
```

- To find the mean and standard deviation use the functions `mean` and `st_deviation`.

[1] Of course, we can't estimate risk on the same data that we used for training!