ACCURACY CONFIDENCE INTERVAL

ESTIMATING & COMPARING THE EFFICACY OF CLASSIFICATION MODELS

Classification and Error Rate

- the error rate using the training set both as training and test set is unavoidable optimistic, with respect to the actual expected error on new data
- the training data might be slightly different from that of test
  - for instance in a bank application the training data to predict the insolvency may regard one region, with the need of extending the result to the entire country
- The data in real problems are usually divided in three subsets
  - training
  - validation, to tune the mining parameters (see the spiral development in CRISP methodology)
  - test, to simulate the error rate on new data

Measuring the confidence range of accuracy of a mining model

- let's suppose that a classifier test has predicted a success rate, namely an accuray, of 75%
- how much this accuray is true for the entire data population ?
  - 75% ± ??? not a single value but a range of accuracies
  - the range of accuracy depends from the size of the test set
  - how much the text set size influnces the accuracy ?
- Let's apply a statistical reasoning

Modeling Classification as a Bernoulli Process

- A classification of N instances can be modelled as Bernoulli process of N independent binary events, e.g. success or error
  - example: toss of a coin
  - if with 100 coin tosses we get 75 heads, which is the probability ***p*** of getting head in next coin toss ? and after 1000 coin tosses ?
  - let's denote N experiments, S successes (num. of correct classifications)
  - ***f*** = S/N success rate (our ACCURACY)
- Confidence Interval
  - given ***f***, may we predict the actual accuray ***p*** of a classification model ?
  - ***p*** is within an interval, with a given probability, namely the confidence
  - N=100 $\Rightarrow$ ***p*** $\in$ [69.1, 80.1] with confidence (i.e. probability) of 80%
  - N=1000 $\Rightarrow$ ***p*** $\in$ [73.2, 76.7] with confidence (i.e. probability) of 80%
    - When N increases, the confidence interval gets smaller

Bernoulli Process (ii)

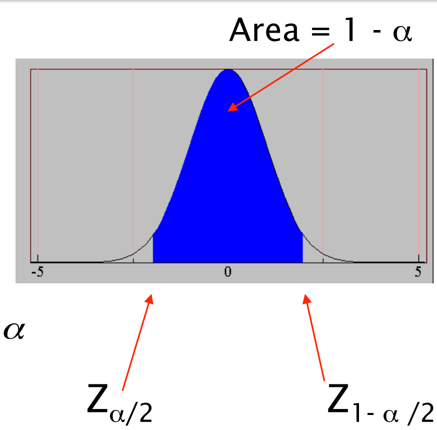
- N experiments: **f** = S/N (**accuracy**)
  - **f** has binomial distribution Bin(N, p) with average **p** and variance **p(1-p)/N**
  - **p** is the actual accuracy we want to estimate
  - for large N value (N greater than 30) the distribution of **f** can be approximated with the normal standard **z distribution**
  - Pr[**-z** ≤ (**f** – **p**) ≤ **z**] = **confidence** (pre-computed in table for unitary standard deviation, see later)
  - Example with confidence of 90%:
    - **z** = 1.65 ➔ Pr[-1.65 ≤ (**f** – **p**) ≤1.65] = 90% --- given **z** let's resolve for **p**

$$Pr\left[-z < \frac{f-p}{\sqrt{p(1-p)/N}} < z\right] = c$$

$$p = \left(f + \frac{z^2}{2N} \pm z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) / \left(1 + \frac{z^2}{N}\right)$$

In Depth Analysis of Confidence Interval

- When in test set N > 30
  - the accuracy approximates the normal standard distribution with average p and variance p(1-p)/N (*acc = f*)



$$P(Z_{\alpha/2} < \frac{acc-p}{\sqrt{p(1-p)/N}} < Z_{1-\alpha/2}) = 1-\alpha$$

- Resolving for *p* we get the confidence Interval:

$$p = \frac{2 \times N \times acc + Z_{\alpha/2}^2 \pm Z_{\alpha/2}\sqrt{Z_{\alpha/2}^2 + 4 \times N \times acc - 4 \times N \times acc^2}}{2(N + Z_{\alpha/2}^2)}$$

Confidence Interval of Accuracy: Example

- Let's consider a model with 80% accuracy evaluated according to a test set of 100 istances:
  - N = 100, acc = 0.8
  - Let's 1-α = 0.95 (95% confidence)
  - From the table we get

    $Z_{\alpha/2} = 1.96$
  - By replacing these values in the preceding formula we get:

1-α	Z
0.99	2.58
0.98	2.33
0.95	1.96
0.90	1.65

N	50	100	500	1000	5000
p min	0.670	0.711	0.763	0.774	0.789
p max	0.888	0.866	0.833	0.824	0.811

Comparing the Accuracy of Two Models

- Given two models M1 ed M2, which is the best ?
  - M1, which has been tested with a data set D1 with cardinality n1, has an error e_1
  - M2, which has been tested with a data set D2 with cardinality n2, has ann error e_2
  - If n1 ed n2 are sufficient large (> 30) **their errors** can be approximated by a Normal distribution with average µ e standard deviation σ:

$$e_1 \sim N(\mu_1, \sigma_1) \qquad e_2 \sim N(\mu_2, \sigma_2)$$

  - The approximated variance is: $\hat{\sigma}_i^2 = \dfrac{e_i(1-e_i)}{n_i}$

Comparing the Accuracy of two Models (ii)

- How can we check if the difference **d** between the two models' accuracies is statistically significant ?
- Lets' **d** = e1 – e2
  - d ~ *N*(d_t,σ_t)   where d_t is the actual difference to estimate
  - the variance σ_t^2 is achieved as follows

$$\sigma_t^2 = \sigma_1^2 + \sigma_2^2 \cong \hat{\sigma}_1^2 + \hat{\sigma}_2^2$$
$$= \frac{e1(1-e1)}{n1} + \frac{e2(1-e2)}{n2}$$

Finally d_t (with confidence 1-α) is

$$d_t = d \pm Z_{\alpha/2} \hat{\sigma}_t$$

Comparing two Models: Example

- Let's M1: n1 = 30, e1 = 0.15
  M2: n2 = 5000, e2 = 0.25
- d = |e2 – e1| = 0.1

$$\hat{\sigma}_d^2 = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

- With confidence 1-α = 0.95, $Z_{\alpha/2}$ = 1.96

$$d_t = 0.100 \pm 1.96 \times \sqrt{0.0043} = 0.100 \pm 0.128$$

=> the interval contains 0 => the difference between the 2 models is **not statistically significant**

Which Confidence Level Makes Significant the Difference between Models ?

- **Let's M1:** n1 = 30, e1 = 0.15 **M2:** n2 = 5000, e2 = 0.25
- d = |e2 – e1| = 0.1

$$\hat{\sigma}_d^2 = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

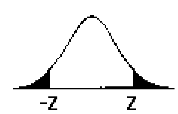
- Which is the confidence threshold to accept the hypothesis that their difference becomes statistically significant ?
- We should determine the value of $\mathbf{Z_{\alpha/2}}$ such that

$$-d < Z_{\alpha/2}\hat{\sigma}_t < d \text{ that is } -\frac{d}{\hat{\sigma}_t} < Z_{\alpha/2} < \frac{d}{\hat{\sigma}_t} \text{ i.e. } 1\text{-}\alpha = P\left(-\frac{d}{\hat{\sigma}_t} < Z_{\alpha/2} < \frac{d}{\hat{\sigma}_t}\right)$$

- Replacing in the example d and σ_t we get $Z_{\alpha/2}$ = ±1.527 ≈ ±1.53
  - that corresponds to α = 0.126, 1-α = 0.874 hence the ***difference becomes significant when the confidence is < 0.874***

Table to Compute the Confidence (1-α) from Z

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	1.000	0.992	0.984	0.976	0.968	0.960	0.952	0.944	0.936	0.928
0,1	0.920	0.912	0.904	0.897	0.889	0.881	0.873	0.865	0.857	0.849
0,2	0.841	0.834	0.826	0.818	0.810	0.803	0.795	0.787	0.779	0.772
0,3	0.764	0.757	0.749	0.741	0.734	0.726	0.719	0.711	0.704	0.697
0,4	0.689	0.682	0.674	0.667	0.660	0.653	0.646	0.638	0.631	0.624
0,5	0.617	0.610	0.603	0.596	0.589	0.582	0.575	0.569	0.562	0.555
0,6	0.549	0.542	0.535	0.529	0.522	0.516	0.509	0.503	0.497	0.490
0,7	0.484	0.478	0.472	0.465	0.459	0.453	0.447	0.441	0.435	0.430
0,8	0.424	0.418	0.412	0.407	0.401	0.395	0.390	0.384	0.379	0.373
0,9	0.368	0.363	0.358	0.352	0.347	0.342	0.337	0.332	0.327	0.322
1,0	0.317	0.312	0.308	0.303	0.298	0.294	0.289	0.285	0.280	0.276
1,1	0.271	0.267	0.263	0.258	0.254	0.250	0.246	0.242	0.238	0.234
1,2	0.230	0.226	0.222	0.219	0.215	0.211	0.208	0.204	0.201	0.197
1,3	0.194	0.190	0.187	0.184	0.180	0.177	0.174	0.171	0.168	0.165
1,4	0.162	0.159	0.156	0.153	0.150	0.147	0.144	0.142	0.139	0.136
1,5	0.134	0.131	0.129	0.126	0.124	0.121	0.119	0.116	0.114	0.112
1,6	0.110	0.107	0.105	0.103	0.101	0.099	0.097	0.095	0.093	0.091
1,7	0.089	0.087	0.085	0.084	0.082	0.080	0.078	0.077	0.075	0.073
1,8	0.072	0.070	0.069	0.067	0.066	0.064	0.063	0.061	0.060	0.059
1,9	0.057	0.056	0.055	0.054	0.052	0.051	0.050	0.049	0.048	0.047
2,0	0.046	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037
2,1	0.036	0.035	0.034	0.033	0.032	0.032	0.031	0.030	0.029	0.029
2,2	0.028	0.027	0.026	0.026	0.025	0.024	0.024	0.023	0.023	0.022
2,3	0.021	0.021	0.020	0.020	0.019	0.019	0.018	0.018	0.017	0.017
2,4	0.016	0.016	0.016	0.015	0.015	0.014	0.014	0.014	0.013	0.013
2,5	0.012	0.012	0.012	0.011	0.011	0.011	0.010	0.010	0.010	0.010
2,6	0.009	0.009	0.009	0.009	0.008	0.008	0.008	0.008	0.007	0.007
2,7	0.007	0.007	0.007	0.006	0.006	0.006	0.006	0.006	0.005	0.005
2,8	0.005	0.005	0.005	0.005	0.005	0.004	0.004	0.004	0.004	0.004
2,9	0.004	0.004	0.004	0.003	0.003	0.003	0.003	0.003	0.003	0.003
3,0	0.003									



Example with Z = ±1.53 choosing the row with Z = 1.53 and column with 0.03 α is 0.126 **Confidence** = 1-α = 0.874

Let's Verify the Confidence Threshold of the Previous Example

- Remind M1: n1 = 30, e1 = 0.1 M2: n2 = 5000, e2 = 0.25
- d = |e2 – e1| = 0.1

$$\hat{\sigma}_d^2 = \frac{0.15(1-0.15)}{30} + \frac{0.25(1-0.25)}{5000} = 0.0043$$

- Setting $Z_{\alpha/2}$ = 1.53, according to 0.874 confidence, we get

$$d_t = 0.100 \pm 1.53 \times \sqrt{0.0043} = 0.100 \pm 0.1003$$

  - => as expected the difference between the 2 models is still not statistically significant because the interval contains zero [-0.0003, 0.2003]
- but with $Z_{\alpha/2}$ = 1.52, corresponding to 0.871 confidence, we get

$$d_t = 0.100 \pm 1.52 \times \sqrt{0.0043} = 0.100 \pm 0.099673$$

  - => the difference is significant as the interval no longer contains ZERO

SLIDE ADDENDUM