Statistical Aspects of Pattern Analysis

John Shawe-Taylor
Computational Statistics and Machine Learning
University College London
jst@cs.ucl.ac.uk

Joint work with Nello Cristianini and Tijl De Bie

September, 2009

STRUCTURE

1. Spurious versus real patterns
2. Pattern significance versus pattern stability
3. Multiple hypothesis testing and the Bonferroni correction
4. Composite hypothesis testing and permutation tests
5. Pattern stability and uniform convergence
6. Rademacher complexity
7. Conclusions

Aims:

- Main motivations plus some of the techniques used
- Some thoughts on why statistics: real vs spurious
- Different approaches: significance vs stability
- Assessing statistical significance and stability
- Multi-hypothesis, multi-pattern scenarios

Patterns: real or spurious

- The DNA of the D7S820 genetic locus:

```
  1   aattttttgta ttttttttag agacggggtt tcaccatgtt ggtcaggctg
 51   actatggagt tattttaagg ttaatatata taaagggtat gatagaacac
101   ttgtcatagt ttagaacgaa ctaacgatag atagatagat agatagatag
151   atagatagat agatagatag atagacagat tgatagtttt tttttatctc
201   actaaatagt ctatagtaaa catttaatta ccaatatttg gtgcaattct
251   gtcaatgagg ataaatgtgg aatcgttata attcttaaga atatatattc
301   cctctgagtt tttgatacct cagattttaa ggcc
```

- This sequence contains 17 repeats of the string `atag`.

- Is this a chance pattern or significant?

Patterns: real or spurious

- In the last 27 presidential elections (all since 1900) Missouri has voted for the winner on 25 occasions, that is 92.6%.
- Is this just a chance occurrence or is something unusual going on?
- Is Missouri able to predict the outcome better than chance?

Patterns: real or spurious

- Rabbi Weissmandel looked at equidistant letter sequences (ELSs) in the Hebrew text of the Bible.
- Witztum, Rips and Rosenberg discovered ELSs that encoded ‘prophesies’ of current events.
- Their paper in Statistical Science claimed that the probability of such patterns occurring by chance was 1 in 60000.
- So is the book of Genesis really prophetic?

Patterns: real or spurious

- Many scientific papers back up their claims by quoting significance levels for results based on a statistical analysis.

- The significance indicates the fraction of papers for which such results fail to hold by chance.

- But all researchers have many projects/papers that they plan to publish, but which fail because the results don't appear significant.

- It would therefore appear that the submitted papers are a sample selected based on the significance of their findings, hence invalidating the significance calculations – does this undermine their validity?

Pattern analysis in science

- The last example highlights core role of the statistics of pattern analysis in scientific research.

- It also suggests that not always completely rigorous in the way in which inferences are made.

- Tutorial aims to revisit statistical inference underpinnings of how research is assessed and to question whether recent developments might not point to new approaches.

- For example Wolfram suggests we should create theories from short programs – such an approach may be rendered more feasible with a statistical learning style of analysis.

Significance vs stability

- Traditionally statisticians have applied significance tests to infer results:
  - the test assumes that the sought result does not hold: the so-called null hypothesis
  - It then works out the probability of data like that observed occurring based on this assumption
  - if this probability is sufficiently small the null hypothesis is refuted and the result proven.
- In Machine Learning statistical learning theory has been interested in the stability of patterns
  - Interesting patterns are sought in the sample driven by particular analysis
  - what is the probability that these patterns are specific to this sample and will not recur in new data?

Significance vs stability

- We will seek to draw parallels between the two approaches:
  - underlying aims of inference are common
  - Similar measures arise in the two cases
  - Scientific status of the assertions is analogous
  - Multi-hypothesis parallel of multi-pattern

- But we also want to contrast the two:
  - how can we verify the assumptions each makes?
  - How do they differ in the way in which they derive new scientific insights?
  - What are the advantages/difficulties associated with each?

Theories of pattern analysis

- Basic approach to view pattern analysis from a statistical viewpoint.

- Aim of any theory is to model real/ artificial phenomena so that we can better understand/ predict/ exploit them.

General statistical considerations

- Statistical models usually begin with an assumption that the data is generated by an underlying distribution P that may only be partially known to the learner.

- If we are trying to classify cancerous tissue, there are two distributions, one for cancerous cells and one for healthy ones, but we may only know that the samples are drawn independently from these distributions.

- If we want to test for the significance of frequent substrings in DNA, we may make assumptions about the way DNA is generated, for example base pairs are generated independently at random.

Significance testing

- For significance testing the assumptions encode the null hypothesis that the experiment hopes to refute.

- Usually the distribution subsumes the processes of the natural/artificial world that we are studying.

- Rather than accessing the distribution directly, statistics often works with a P generated ‘training sample’ or ‘training set’ X typically with a number of components

$$X = \{x_1, \ldots, x_m\}$$

- Significance testing assesses the likelihood that ‘such data’ arises given the null hypothesis.

Spurious versus real patterns

- Since we are only assessing significance/stability based on a random sample, we can only make probabilistic assertions.

- We might be very unlucky with the sample and be misled.

- For example if we hypothesise that most cars are sports cars, and observe a random sample driving along a road, by chance half might be sports cars.

- Clearly, the bigger the sample and the stronger the hypothesis (e.g. 90% of cars are sports cars), the less likely we are of being misled.

Summary

- Motivation of statistical analysis: assessing if patterns are chance or not.

- Examples illustrate the range of data and fundamental role in scientific research.

- Two styles of analysis: significance testing vs assessing stability

- Perhaps a moment to reassess how we verify scientific results to verify we're making the best and right use of the data and check we're not missing a new trick.

STRUCTURE

1. Spurious versus real patterns
2. Pattern significance versus pattern stability
3. Multiple hypothesis testing and the Bonferroni correction
4. Composite hypothesis testing and permutation tests
5. Pattern stability and uniform convergence
6. Rademacher complexity
7. Conclusions

Significance testing

- For significance testing the probability of being misled is known as the p-value:

  $$p = P(\pi(\underline{X}) \geq \pi(X))$$

  where $\pi(X)$ is some measure of 'strength' of the pattern π also known as a pattern function:

  $$\pi : X \longmapsto \mathbb{R}$$

- Note the probability is over an independent random sample $\underline{X}$ and so p is the probability that we get at least as strong a pattern as observed.

Example of significance testing

- Recall the state of Missouri has voted for the winner 25 our of the last 27 USA presidential elections (since 1900) – strange or not?

- Let's build a model of the probabilities and test for significance.

- Data is binary vector for each state counting number of winning votes of the $n = 27$ elections

$$X = (x_1, \ldots, x_n) \in \{0, 1\}^n$$

- Our pattern function counts fraction of winning votes:

$$\pi(X_{\text{Missouri}}) = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{25}{27} = 0.926$$

Missouri example cont

- We now need to specify the null hypothesis as the default probability distribution:

- We assume a Bernouilli probability distribution randomly deciding to vote for the winner with probability $q = 0.734$, since this is the average no of votes for the winner over all states. So

$$\begin{aligned} p &= P(\pi(\underline{X}) \geq \pi(X)) \\ &= \sum_{i=25}^{27} P\left(\pi(\underline{X}) = \frac{i}{n}\right) \\ &= \sum_{i=25}^{27} \binom{n}{i} 0.734^{i}(1-0.734)^{n-i} \\ &= 0.014 \end{aligned}$$

Significance level

- Normally a threshold δ is specified before the test is made and the null hypothesis is rejected if $p < \delta$.

- Typical levels for scientific tests are 5%, 2% or 1%: this effectively specifies the level of mistakes we are prepared to tolerate.

- The Missouri results appear to be prophetic at the 2% significance level.

- More precisely the chances of Missouri getting it right that often is less than 2% if we assume all states equally prophetic and independence of results.

Pattern stability

- The aim of pattern stability analysis is to make predictions about future pattern strengths.

- If we observe a certain pattern strength and make perhaps milder assumptions about the distribution, can we infer its expected strength in a new sample.

- Naturally we again have the caveat that we may have been misled.

- For example if we observe the fraction of sports cars over a short period we may wish to conclude a likely lower bound on their real frequency.

Pattern stability

- The strength $\overline{\pi}$ of a pattern π is best measured by its expected value over a randomly generated sample:
$$\overline{\pi} = \mathbb{E}[\pi(\underline{X})]$$
- Given an observation, we would like to give an interval within which $\overline{\pi}$ must fall unless we have been badly misled.
- Since we are only assessing stability based on a random sample, again we can only make probabilistic assertions.
- For example if we hypothesise that the true mean is close to that of the sample it could be a very unrepresentative sample.

Missouri example

- We can relax the assumption about knowing a state's probability of voting for the winner.

- Now we seek the interval within which Missouri's probability will fall with high confidence.

- If we specify the confidence $\delta = 0.05$, then what we seek are the Bernouilli probabilities θ such that the null hypothesis would not be rejected, that is

$$\sum_{i=25}^{27} \binom{n}{i} \theta^i (1-\theta)^{n-i} \geq 0.05$$

- Since $\theta = 0.785$ makes the above an equality, and $\theta = 1$ satisfies the inequality the interval is $[0.785, 1]$

Missouri stability

- We conclude that with confidence at least $1 - \delta = 0.95$ the expected probability with which Missouri picks the winning state is at least 0.785 – again making the assumptions that the decisions are independent Bernouilli variables.

Concentration inequalities

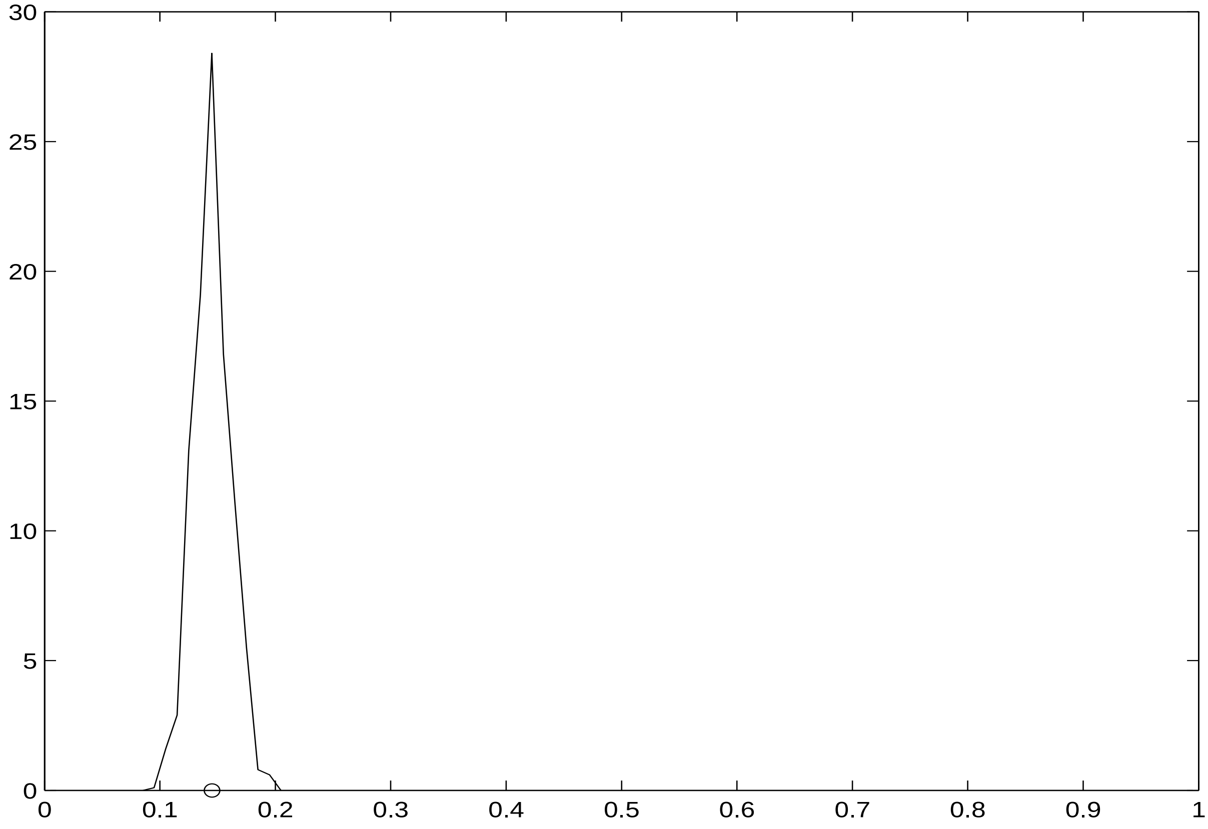
- Statistical Learning is concerned with the reliability or stability of inferences made from a random sample.

- Random variables with this property have been a subject of ongoing interest to probabilists and statisticians.

- For a random variable this corresponds to the distribution clustering tightly around its mean, referred to as concentration.

- For concentrated variables the mean gives with high probability a good estimate of a random draw.

Error distribution: dataset size: 342



Concentration inequalities cont.

- As an example consider the mean of a sample of m 1-dimensional random variables $X_1, \ldots, X_m$:

$$S_m = \frac{1}{m} \sum_{i=1}^{m} X_i.$$

- Hoeffding's inequality states that if $X_i \in [a_i, b_i]$

$$P\{|S_m - \mathbb{E}[S_m]| \geq \epsilon\} \leq 2 \exp\left(-\frac{2m^2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right)$$

Note how the probability falls off exponentially with the distance from the mean and with the number of variables.

McDiarmid’s inequality

Theorem 1. *Let $X_1, \ldots, X_n$ be independent random variables taking values in a set A, and assume that $f : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,\hat{x}_i \in A} |f(x_1, \ldots, x_n) - f(x_1, \ldots, \hat{x}_i, x_{i+1}, \ldots, x_n)| \le c_i,$$

for $1 \le i \le n$. Then for all $\epsilon > 0$,

$$P\left\{f\left(X_1, \ldots, X_n\right) - \mathbb{E}f\left(X_1, \ldots, X_n\right) \ge \epsilon\right\} \le \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

- Hoeffding is a special case when $f(x_1, \ldots, x_n) = S_n$

Using McDiarmid

- By setting the right hand side equal to δ, we can always invert McDiarmid to get a high confidence bound: with probability at least $1 - \delta$

$$f\left(X_1, \ldots, X_n\right) < \mathbb{E} f\left(X_1, \ldots, X_n\right) + \sqrt{\frac{\sum_{i=1}^{n} c_i^2}{2} \log \frac{1}{\delta}}$$

- If $c_i = c/n$ for each i this reduces to

$$f\left(X_1, \ldots, X_n\right) < \mathbb{E} f\left(X_1, \ldots, X_n\right) + \sqrt{\frac{c^2}{2n} \log \frac{1}{\delta}}$$

Application to Missouri

- Replacing one election can only change the rate of success by $c_i = 1/n$, so with probability at least $1 - \delta$

$$\begin{aligned} \pi(\underline{X}) &\geq \mathbb{E}[\pi(\underline{X})] + \sqrt{\frac{1}{2n} \log \frac{1}{\delta}} \\ \Rightarrow \quad \mathbb{E}[\pi(\underline{X})] &\geq 0.764 \end{aligned}$$

  if $\delta = 0.05$

- Note this is wider interval than $[0.785, 1]$ obtained before – but then did exact computation. In general won't be able to do this.

Summary

- Considered pattern significance testing and contrasted with pattern stability.
- In case of stability introduced concentration inequalities as a useful theoretical tool.
- Simple examples seem to suggest that Missouri is prophetic? Need to move to multiple hypothesis testing to see why this isn't the case.

STRUCTURE

1. Spurious versus real patterns
2. Pattern significance versus pattern stability
3. Multiple hypothesis testing and the Bonferroni correction
4. Composite hypothesis testing and permutation tests
5. Pattern stability and uniform convergence
6. Rademacher complexity
7. Conclusions

Missouri revisited

- There has been an underlying flaw in our discussions about Missouri: it was selected precisely because it had such a good prediction record.

- This choice biases any calculations since in a sample of 45 states (in the union since 1900), it is much more likely to observe one that has a good hit rate just by chance.

- To be fair since we have no a priori reason to expect one state to be better than another, we should consider one pattern function for each state:

$$\pi_1, \ldots, \pi_{45}$$

  and test for each.

Multiple hypotheses

- We need to downweight the probability of being misled.

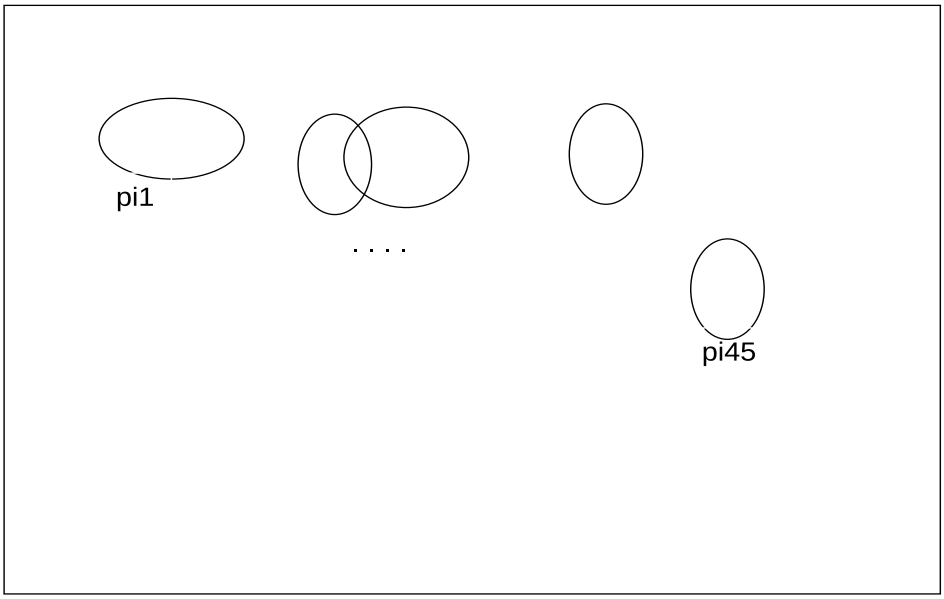
- Recall that significance is probability that ‘such data’ could have arisen by chance.

- Now we have several ways that significance might arise:

$$p = P(\exists i \in \{1, \ldots, 45\} : \pi_i(\underline{X}) \geq \pi_i(X))$$

- So the misleading $\underline{X}$s are those that cause any of the inequalities to hold.

In Picture



- Must ensure that total probability is less than δ.
- There may be overlaps but union bound ignores these by taking sum of probabilities.

Missouri

- If we make each less than $\delta/45$ then total probability less than δ
- So to achieve significance at 0.05 need value c of $\pi_i(X)$ such that

$$\begin{aligned} \text{for } \delta = 0.05,\ P(\pi_i(\underline{X}) \geq c) &\leq \delta/45 = 0.0019 \\ P(\pi_i(\underline{X}) \geq 26/27) &= 0.0026 \\ P(\pi_i(\underline{X}) \geq 27/27) &= 0.0002 \end{aligned}$$

- So the value is $c = 27/27$.
- Hence $c = 0.926$ is not significant.

Bonferroni correction

- The adjustment of $\delta \rightarrow \delta/|\Pi|$ for performing significance tests over a set Π of patterns is known as the Bonferroni correction.

- The correction doesn't have to be uniform: can use $\delta \rightarrow \delta q_i$ for pattern i provided

$$\sum_{i\in\Pi} q_i \leq 1$$

- The choice of q_i encodes our prior belief about which patterns are more likely to occur.

- We chose the uniform correction as we have no reason to expect one state to be more predictive than another.

Tandem repeats example

- Recall that there was a repeat of a length 4 string 17 times.

- If we consider all strings of length 2 to 4, there are $4^4 + 4^3 + 4^2 = 336$ strings.

- Hence, $\delta \rightarrow \delta/336$, and since

$$p = P(\exists \pi \in \Pi : \pi(\underline{X}) > 12) < 0.01$$

the tandem repeats are significant at the 1% level assuming iid generation.

Tandem repeats example

- In reality we may not know the length of the strings that might turn out to be significant, but choosing a uniform division of δ cuts the value for small strings too much.

- Natural to stratify the division by the length: give $\delta/2^\ell$ to strings of length ℓ and then subdivide this uniformly over the strings of length ℓ.

- Corresponds to $\delta \to \delta(\pi) = \delta/(2|\Sigma|)^{\ell_\pi}$ where Σ is the alphabet and ℓ_π the length of pattern π.

- This encodes a prior belief that the frequent strings will not be very long.

Stratification

- Following the capacity functional notation, if we specify an upper bound on the capacity functional, i.e. a lower bound on the length of the string, and then maximise the strength:

$$\begin{array}{ll} \max\limits_{\pi} & \pi(X) \\ \textbf{s.t.} & S(\pi) \geq S_{\min} \end{array}$$

  the solution will always have length $S_{\min}$ so no need to stratify δ.

- But need to choose $S_{\min}$ ahead of time, or do multiple tests over the different values.

Linking with algorithms

- Consider the algorithm that places all of the substrings of a given length at the root of a trie (each node's children indexed by the alphabet Σ).

- the strings are progressively moved down the tree into the branch indexed by their next character.

- any node with no strings is not created.

- Counts at the leaves (at depth $S_{\min}$) reveal the frequent strings.

Optimising the algorithm

- We can precompute the count required at a leaf to ensure significance.

- Any internal nodes having fewer than this number of substrings reaching them can be pruned.

- We can also view the division of δ as happening as we traverse the tree.

- This gives a natural way to consider different divisions depending on whether we want to test for significance at the internal nodes, etc.

- There does not appear to be a simple way to take into account the overlap between bad events.

STRUCTURE

1. Spurious versus real patterns
2. Pattern significance versus pattern stability
3. Multiple hypothesis testing and the Bonferroni correction
4. Composite hypothesis testing and permutation tests
5. Pattern stability and uniform convergence
6. Rademacher complexity
7. Conclusions

Composite hypothesis testing

- In some cases, the null hypothesis is not defined exactly as a single probability density P, but rather as a set $\mathcal{P}$ of densities

- Composite hypothesis test: A hypothesis test with significance level δ is called conservative with respect to a composite null hypothesis $\mathcal{P}$, if it satisfies:

$$\forall P \in \mathcal{P}, \qquad P(\pi(\underline{X}) \geq t_\delta) \leq \delta.$$

  I.e. for each of the probability distributions $P \in \mathcal{P}$, rejecting the null hypothesis while it holds true is at most equal to the significance level δ.

Symmetric Composite hypotheses

- Often, the null hypothesis is defined in terms of symmetries

- The prophetic state: We can assume as a null hypothesis that permutations of the votes among the states (within each year) are equally likely:

  $$P\left(\underline{X}\right) = P\left(T\left(\underline{X}\right)\right)$$

  where $T \in \mathfrak{T}$, here the set of permutations of the votes over the states (for each year separately).

Symmetric hypothesis testing

- Given that the null hypothesis is defined in terms of transformation invariants $T \in \mathcal{T}$, the p-value can be computed as:

  $$p = \frac{|T \in \mathcal{T} : \pi\left(T\left(X\right)\right) \geq \pi\left(X\right)|}{|\mathcal{T}|},$$

  the proportion of transformations T for which $\pi\left(T\left(X\right)\right) \geq \pi\left(X\right)$

- As if generating new random data $T\left(X\right)$ from the data X, drawn from the same (only partially specified) distribution P

- In practice, only a randomly sampled subset from $\mathcal{T}$ is used... (computational reasons)

The DNA example

- Tandem repeats in ‘junk’-DNA: Assuming the DNA is really junk, it makes sense to assume all sequence permutations are equally likely: the null hypothesis.

- Check this by permuting 10,000 times and computing the pattern strength on each permuted string. The p-value = the proportion of times that $\pi\left(T\left(X\right)\right) \geq 17$,

- here $< \frac{1}{10{,}000}$.

STRUCTURE

1. Spurious versus real patterns

2. Pattern significance versus pattern stability

3. Multiple hypothesis testing and the Bonferroni correction

4. Composite hypothesis testing and permutation tests

5. Pattern stability and uniform convergence

6. Rademacher complexity

7. Conclusions

Multi-patterns for pattern stability

- We have seen how testing multiple hypotheses leads to the Bonferroni correction based on the union bound.

- If we have a finite set of pattern functions we can apply exactly the same idea to bound the probability that the estimates of the means of each lie in a set of intervals.

- In this context this is known as uniform convergence, since we effectively ask that we have statistical convergence uniformly over all of the patterns.

Probability of being misled in classification

- Basic approach is again to bound the probability of being misled and set this equal to δ – take classification example.

- What is the chance of being misled by a single bad function f, i.e. training error $\mathrm{err}_S(f) = 0$, while true error is bad $\mathrm{err}(f) > \epsilon$?

$$\begin{aligned} P_S\left\{\mathrm{err}_S(f) = 0, \mathrm{err}(f) > \epsilon\right\} &= (1 - \mathrm{err}(f))^m \\ &\leq (1-\epsilon)^m \\ &\leq \exp(-\epsilon m). \end{aligned}$$

so that choosing $\epsilon = \ln(1/t)/m$ ensures probability less than t.

Finite or Countable function classes

If we now consider a function class

$$\mathcal{F} = \{f_1, f_2, \ldots, f_n, \ldots\}$$

and make the probability of being misled by f_n less than $q_n\delta$ with

$$\sum_{n=1}^{\infty} q_n \leq 1,$$

then the probability of being misled by one of the functions is bounded by

$$P_S \left\{ \exists f_n \colon \mathrm{err}_S(f_n) = 0, \mathrm{err}(f_n) > \frac{1}{m} \ln\left(\frac{1}{q_n\delta}\right) \right\} \leq \delta.$$

This again uses the union bound.

Finite or Countable function classes result

- The bound translates into a theorem: given $\mathcal{F}$ and q, with probability at least $1-\delta$ over random m samples the generalisation error of a function $f_n \in \mathcal{F}$ with zero training error is bounded by

$$\mathrm{err}(f_n) \leq \frac{1}{m}\left(\ln\left(\frac{1}{q_n}\right) + \ln\left(\frac{1}{\delta}\right)\right)$$

- We can think of the term $\ln\left(\frac{1}{q_n}\right)$ as the complexity / description length of the function f_n.

Some comments on the result

- Note that we must put a prior weight on the functions. If the functions are drawn at random according to a distribution p_n, the expected generalisation will be minimal if we choose our prior $q = p$.

- Interestingly if true distribution with which classifiers arise is p then using q makes the average bound worse by

$$\frac{1}{m}\mathrm{KL}(p\|q).$$

- This is the starting point of the PAC-Bayes analysis.

Statistical learning theory

- SLT has developed methods of proving uniform convergence over sets that have uncountably many elements where the union bound must fail.

- Hence, the techniques take into account the overlap between different patterns in order to beat the union bound.

- Earliest technique used the VC dimension, but more refined methods are based on Rademacher complexity.

STRUCTURE

1. Spurious versus real patterns
2. Pattern significance versus pattern stability
3. Multiple hypothesis testing and the Bonferroni correction
4. Composite hypothesis testing and permutation tests
5. Pattern stability and uniform convergence
6. Rademacher complexity
7. Conclusions

Rademacher complexity

- Rademacher complexity measures the complexity of a function class by asking how well it can correlate with noise:

$$R_m(\mathcal{F}) = \mathbb{E}_{S\sigma}\left[\sup_{f\in\mathcal{F}} \frac{2}{m}\sum_{i=1}^{m} \sigma_i f\left(\mathbf{z}_i\right)\right].$$

  is known as the Rademacher complexity of the function class $\mathcal{F}$, where σ_i are uniformly random ± 1 variables and expectation is also over random iid sample S.

Rademacher proof beginnings

For a fixed $f \in \mathcal{F}$ we have

$$\mathbb{E}\left[f(\mathbf{z})\right] \leq \hat{\mathbb{E}}\left[f(\mathbf{z})\right] + \sup_{h \in \mathcal{F}} \left(\mathbb{E}[h] - \hat{\mathbb{E}}[h]\right).$$

where $\mathcal{F}$ is a class of functions mapping from Z to $[0, 1]$ and $\hat{\mathbb{E}}$ denotes the sample average.

We must bound the size of the second term. First apply McDiarmid's inequality to obtain ($c_i = 1/m$ for all i) with probability at least $1 - \delta$:

$$\sup_{h \in \mathcal{F}} \left(\mathbb{E}[h] - \hat{\mathbb{E}}[h]\right) \leq \mathbb{E}_S \left[\sup_{h \in \mathcal{F}} \left(\mathbb{E}[h] - \hat{\mathbb{E}}[h]\right)\right] + \sqrt{\frac{\ln(1/\delta)}{2m}}.$$

Deriving double sample result

- We can now move to the ghost sample by simply observing that $\mathbb{E}[h] = \mathbb{E}_{\tilde{S}}\left[\hat{\mathbb{E}}[h]\right]$:

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{F}}\left(\mathbb{E}[h]-\hat{\mathbb{E}}[h]\right)\right] =$$

$$\mathbb{E}_S\left[\sup_{h\in\mathcal{F}}\mathbb{E}_{\tilde{S}}\left[\frac{1}{m}\sum_{i=1}^{m}h(\tilde{\mathbf{z}}_i)-\frac{1}{m}\sum_{i=1}^{m}h(\mathbf{z}_i)\middle| S\right]\right]$$

Deriving double sample result cont.

Since the sup of an expectation is less than or equal to the expectation of the sup (we can make the choice to optimise for each $\tilde{S}$) we have

$$\mathbb{E}_S \left[\sup_{h \in \mathfrak{F}} \left(\mathbb{E}[h] - \hat{\mathbb{E}}[h] \right) \right] \leq \mathbb{E}_S \mathbb{E}_{\tilde{S}} \left[\sup_{h \in \mathfrak{F}} \frac{1}{m} \sum_{i=1}^{m} \left(h(\tilde{\mathbf{z}}_i) - h(\mathbf{z}_i) \right) \right]$$

Adding symmetrisation

Here symmetrisation is again just swapping corresponding elements – but we can write this as multiplication by a variable σ_i which takes values ± 1 with equal probability:

$$
\begin{aligned}
\mathbb{E}_S \left[\sup_{h \in \mathcal{F}} \left(\mathbb{E}[h] - \hat{\mathbb{E}}[h] \right) \right] \leq & \\
\leq & \quad \mathbb{E}_{\sigma S \tilde{S}} \left[\sup_{h \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \left(h(\tilde{\mathbf{z}}_i) - h(\mathbf{z}_i) \right) \right] \\
\leq & \quad 2\mathbb{E}_{\sigma S} \left[\sup_{h \in \mathcal{F}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h(\mathbf{z}_i) \right] \\
= & \quad R_m \left(\mathcal{F} \right),
\end{aligned}
$$

Main Rademacher theorem

Putting the pieces together gives the main theorem of Rademacher complexity: with probability at least $1-\delta$ over random samples S of size m, every $f \in \mathcal{F}$ satisfies

$$\mathbb{E}\left[f(\mathbf{z})\right] \leq \hat{\mathbb{E}}\left[f(\mathbf{z})\right] + R_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

- Note that Rademacher complexity gives the expected value of the maximal correlation with random noise – a very natural measure of capacity.

- Note that the Rademacher complexity is distribution dependent since it involves an expectation over the choice of sample – this might seem hard to compute.

Empirical Rademacher theorem

- Since the empirical Rademacher complexity

$$\hat{R}_m(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{2}{m} \sum_{i=1}^{m} \sigma_i f\left(\mathbf{z}_i\right) \middle| \mathbf{z}_1, \ldots, \mathbf{z}_m \right]$$

is concentrated, we can make a further application of McDiarmid to obtain with probability at least $1 - \delta$

$$\mathbb{E}_\mathcal{D}\left[f(\mathbf{z})\right] \leq \hat{\mathbb{E}}\left[f(\mathbf{z})\right] + \hat{R}_m(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}.$$

- For class $\mathcal{F}_B$ of linear functions with norm B we can bound RC:

Rademacher complexity of $\mathcal{F}_B$

The following derivation gives the result

$$\begin{aligned}
\hat{R}_m(\mathcal{F}_B) &= \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_B}\frac{2}{m}\sum_{i=1}^m \sigma_i f\left(\mathbf{x}_i\right)\right] \\
&= \mathbb{E}_\sigma\left[\sup_{\|\mathbf{w}\|\leq B}\left\langle \mathbf{w}, \frac{2}{m}\sum_{i=1}^m \sigma_i \phi\left(\mathbf{x}_i\right)\right\rangle\right] \\
&\leq \frac{2B}{m}\mathbb{E}_\sigma\left[\left\|\sum_{i=1}^m \sigma_i\phi(\mathbf{x}_i)\right\|\right] \\
&= \frac{2B}{m}\mathbb{E}_\sigma\left[\left(\left\langle\sum_{i=1}^m \sigma_i\phi(\mathbf{x}_i), \sum_{j=1}^m \sigma_j\phi(\mathbf{x}_j)\right\rangle\right)^{1/2}\right]
\end{aligned}$$

$$\leq \frac{2B}{m}\left(\mathbb{E}_\sigma\left[\sum_{i,j=1}^m \sigma_i\sigma_j\kappa(\mathbf{x}_i,\mathbf{x}_j)\right]\right)^{1/2} = \frac{2B}{m}\sqrt{\sum_{i=1}^m \kappa(\mathbf{x}_i,\mathbf{x}_i)}$$

SVM bound

- Assembling the result we obtain that with probability at least $1 - \delta$ over the generation of the training data:

$$\begin{aligned} P(y \neq \mathrm{sgn}(g(\mathbf{x}))) &= \mathbb{E}\left[\mathcal{H}(-yg(\mathbf{x}))\right] \\ &\leq \frac{1}{m\gamma}\sum_{i=1}^{m}\xi_i + \frac{4}{m\gamma}\sqrt{\sum_{i=1}^{m}\kappa(\mathbf{x}_i, \mathbf{x}_i)} + 3\sqrt{\frac{\ln(2/\delta)}{2m}} \end{aligned}$$

- Note that for the Gaussian kernel this reduces to

$$P(y \neq \mathrm{sgn}(g(\mathbf{x}))) \leq \frac{1}{m\gamma}\sum_{i=1}^{m}\xi_i + \frac{4}{\sqrt{m}\gamma} + 3\sqrt{\frac{\ln(2/\delta)}{2m}}$$

Kernel PCA

- the projection of a new point into the space spanned by the i-th eigenvector of the correlation matrix

  $$C(S) = \frac{1}{m}\sum_{i=1}^{m}\phi(\mathbf{x}_i)\phi(\mathbf{x}_i)'$$

  of a sample S can be computed as

  $$P_{\mathbf{u}_i}(\phi(\mathbf{x})) = \hat{\lambda}_i^{-1/2}\sum_{j=1}^{m}\mathbf{v}_{ij}\kappa(\mathbf{x},\mathbf{x}_j),$$

  where $(\mathbf{v}_{ij})_{j=1}^{m}$, $\hat{\lambda}_i$ are the i-th eigenvector and eigenvalue of the kernel matrix $K(S)$.

Kernel PCA cont.

- Hence we can perform PCA in a kernel defined feature space in the orthonormal basis given by the eigen-vectors of $C(S)$.

- PCA standard technique applied for low dimensional feature spaces – no guarantee that it is always sensible to use the approach in the high-dimensional feature spaces typical of kernel methods.

Analysis of Kernel PCA

- SLT should be able to highlight the critical elements which affect the quality of the kernel PCA.

- Consider performing PCA on a randomly drawn training set S of size m in the feature space defined by a kernel $\kappa(\mathbf{x}, \mathbf{z})$ and project new data onto the space $\hat{V}$ spanned by the first k eigenvectors

Statistical analysis of PCA

- with probability greater than $1 - \delta$ over the generation of the sample S the expected squared residual is bounded by

$$\mathbb{E}\left[\|P_{\hat{V}}^{\perp}(\phi(\mathbf{x}))\|^2\right] \leq \frac{1}{m}\sum_{i=k+1}^{m}\hat{\lambda}_i(S)$$

$$+\frac{1+\sqrt{k}}{\sqrt{m}}\sqrt{\frac{2}{m}\sum_{i=1}^{m}\kappa(\mathbf{x}_i,\mathbf{x}_i)^2} + R^2\sqrt{\frac{18}{m}\ln\left(\frac{2}{\delta}\right)},$$

where the support of the distribution is in a ball of radius R in the feature space.

Outline of proof

- Let $X = U\Sigma V'$ be the singular value decomposition of the sample matrix X in the feature space. The projection norm is then given by

  $$\hat{f}(\mathbf{x}) = \|P_{\hat{V}}(\phi(\mathbf{x}))\|^2 = \phi(\mathbf{x})' U_k U_k' \phi(\mathbf{x}),$$

  where U_k is the matrix containing the first k columns of U.

Outline of proof

- Hence we can write

$$\|P_{\hat{V}}(\phi(\mathbf{x}))\|^2 = \sum_{ij=1}^{N_F} w_{ij}\phi(\mathbf{x})_i\phi(\mathbf{x})_j = \sum_{ij=1}^{N_F} w_{ij}\hat{\phi}(\mathbf{x})_{ij},$$

  where $\hat{\phi}$ is the projection mapping into the feature space $\hat{F}$ consisting of all pairs of F features and $w_{ij} = (U_k U_k')_{ij}$.

Feature space construction

- The standard polynomial construction gives

$$\begin{aligned}
\hat{\kappa}(\mathbf{x},\mathbf{z}) &= \kappa(\mathbf{x},\mathbf{z})^2 = \left(\sum_{i=1}^{N_F} \phi(\mathbf{x})_i \phi(\mathbf{z})_i\right)^2 \\
&= \sum_{i,j=1}^{N_F} \phi(\mathbf{x})_i \phi(\mathbf{z})_i \phi(\mathbf{x})_j \phi(\mathbf{z})_j \\
&= \left\langle \hat{\phi}(\mathbf{x}), \hat{\phi}(\mathbf{z}) \right\rangle_{\hat{F}} .
\end{aligned}$$

Feature space construction cont.

- The norm of $\hat{f}$ satisfies (note that $\| \cdot \|_F$ denotes the Frobenius norm)

$$
\begin{aligned}
\|\hat{f}\|^2 &= \sum_{i,j=1}^{N_F} \alpha_{ij}^2 = \|U_k U_k'\|_F^2 \\
&= \left\langle \sum_{i=1}^{k} \mathbf{u}_i \mathbf{u}_i', \sum_{j=1}^{k} \mathbf{u}_j \mathbf{u}_j' \right\rangle_F = \sum_{i,j=1}^{k} (\mathbf{u}_i' \mathbf{u}_j)^2 = k
\end{aligned}
$$

Applying Rademacher complexity

- We consider the function class $\hat{\mathcal{F}}_{\sqrt{k}}$ with respect to the kernel

$$\hat{\kappa}(\mathbf{x}, \mathbf{z}) = \kappa(\mathbf{x}, \mathbf{z})^2,$$

augmenting the corresponding primal weight vectors with one further dimension while augmenting the corresponding input vectors with a feature

$$\begin{aligned} \|\phi(\mathbf{x}))\|^2 k^{-0.25} &= \kappa(\mathbf{x}, \mathbf{x}) k^{-0.25} = k^{-0.25}\sqrt{\hat{\kappa}(\mathbf{x}, \mathbf{x})} \\ &= \|\hat{\phi}(\mathbf{x}))\| k^{-0.25} \end{aligned}$$

Applying Rademacher complexity cont.

- We now apply the Rademacher theorem to the class

$$
\begin{aligned}
\hat{F} \quad &= \quad \Big\{ f_\ell : (\hat{\phi}(\mathbf{x}), \|\hat{\phi}(\mathbf{x}))\| k^{-0.25}) \\
&\qquad \mapsto (\|\hat{\phi}(\mathbf{x}))\| - f(\hat{\phi}(\mathbf{x})))R^{-2} \mid f \in \hat{\mathcal{F}}_{\sqrt{k}} \cap \mathcal{P} \Big\} \\
&\subseteq \quad R^{-2}\hat{\mathcal{F}}'_{\sqrt{k+\sqrt{k}}},
\end{aligned}
$$

Hypothesis testing using kernel spaces

- Work by Gretton et al. for testing if two sets of data are drawn from different distributions.

- Consider a kernel defined function space:

$$\mathcal{F} = \{\mathbf{x} \mapsto \langle \mathbf{w}, \phi(\mathbf{x}) \rangle : \|\mathbf{w}\| \leq 1\}$$

- Define maximum discrepancy pattern function over two equal sized samples S and $\tilde{S}$:

$$m(S, \tilde{S}) = \sup_{f \in \mathcal{F}} \left(\hat{\mathbb{E}}_S[f] - \hat{\mathbb{E}}_{\tilde{S}}[f] \right)$$

- Note that we can define a test for each $f \in \mathcal{F}$, so this can be viewed as a multiple test over infinitely many hypotheses.

Composite hypothesis and permutation testing

- Null hypothesis is that two samples are drawn i.i.d. from the same distribution:

- i.e. the composite hypothesis test – data generated i.i.d. from any (fixed) distribution with support in the unit ball.

- Hence, can use permutation testing since all permutations of the data are equally likely

- actually use the same swapping permutations specified by the Rademacher variables to estimate expected value:

$$\mathbb{E}_\sigma[m(S,\tilde{S})] = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i(f(\mathbf{x}_i)-f(\tilde{\mathbf{x}}_i))\right)\right]$$

Bounding expected value

- Hence, we have

$$\begin{aligned}\mathbb{E}_{S\tilde{S}}\mathbb{E}_{\sigma}[m(S,\tilde{S})] &\leq 2\mathbb{E}_{\sigma S}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(\mathbf{x}_i)\right)\right] \\ &\leq \frac{2}{m}\sqrt{\mathrm{Tr}(K_S)}\end{aligned}$$

- Applying McDiarmid to bound the probability that we are far from the mean over $S, \tilde{S}$, we have with probability at least $1-\delta$

$$m(S,\tilde{S}) \leq \frac{2}{m}\sqrt{\mathrm{Tr}(K_S)} + \sqrt{\frac{2}{m}\log\frac{1}{\delta}}$$

Applying the test

- need to compute $m(S, \tilde{S})$ on the two samples:

$$\begin{aligned} m(S, \tilde{S}) &= \frac{1}{m} \sup_{\mathbf{w}:\|\mathbf{w}\| \leq 1} \sum_{i=1}^{m} \left(\langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle - \langle \mathbf{w}, \phi(\tilde{\mathbf{x}}_i) \rangle \right) \\ &= \frac{1}{m} \sup_{\mathbf{w}:\|\mathbf{w}\| \leq 1} \left\langle \mathbf{w}, \sum_{i=1}^{m} (\phi(\mathbf{x}_i) - \phi(\tilde{\mathbf{x}}_i)) \right\rangle \\ &= \frac{1}{m} \left\| \sum_{i=1}^{m} (\phi(\mathbf{x}_i) - \phi(\tilde{\mathbf{x}}_i)) \right\| = \frac{1}{m} \sqrt{\mathbf{y}' K_{S\tilde{S}} \mathbf{y}} \end{aligned}$$

where $\mathbf{y}$ is the vector with 1s for S and -1s for $\tilde{S}$

- Hence, test for significance at level δ is:

$$\sqrt{\mathbf{y}' K_{S\tilde{S}} \mathbf{y}} > 2\sqrt{\mathrm{Tr}(K_S)} + \sqrt{2m \log \frac{1}{\delta}}$$

Conclusions

- Statistical analysis of patterns seen as crucial in assessing if real or spurious.
- General framework includes both significance testing and stability analysis.
- Multiple hypothesis testing can be effected using the Bonferroni correction based on the union bound.
- Methods from SLT can also be used in hypothesis testing.
- Example given of detecting if two data samples are drawn from different distributions.