Pattern Recognition (PR)

Prof. Dr.-Ing. Andreas Maier
Pattern Recognition Lab (CS 5), Friedrich-Alexander-Universität Erlangen-Nürnberg
Winter Term 2020/21

This is a printable version of the slides of the lecture

Pattern Recognition (PR)
Winter term 2020/21
Friedrich-Alexander University of Erlangen-Nuremberg.

Erlangen, October 28, 2020
Prof. Dr.-Ing. Andreas Maier

Model Assessment

No Free Lunch

- In the past lectures, we have come across many learning algorithms and classification techniques.
- They have properties such as
  - low computational complexity
  - incorporation of prior knowledge
  - linearity / non-linearity
  - optimality with respect to certain cost functions, etc.
- Some compute smooth decision boundaries, some compute rather non-smooth decision boundaries.

We really have to ask:

Are there any reasons to favor one algorithm over another?

No Free Lunch (cont.)

Theorem

Given a cost function $f \in \mathscr{F}$, an algorithm A and costs c_m for a specific sample that is iterated on m times.

The performance of an algorithm is the conditional probability $P(c_m | f, m, A)$.

The No Free Lunch Theorem states that for any two algorithms A_1 and A_2:

$$\sum_f P(c_m | f, m, A_1) = \sum_f P(c_m | f, m, A_2)$$

No Free Lunch (cont.)

Consequences for classification methods:

- If no prior assumptions about the problem are made, there is **NO** overall superior or inferior classification method!
- We should be skeptical regarding studies that demonstrate the overall superiority of a particular method.
- We have to focus on the aspects that matter most for the classification problem:
  - prior information
  - data distribution
  - amount of training data
  - cost functions

Off-Training Set Error

Off-training set error:

- Specifies the error on samples that are not contained within the training set.
- For large training data sets, the off-training set is necessarily small.
- Used to compare general classification performance of algorithms.

Off-Training Set Error (cont.)

- Consider a two-class problem with training data set $\mathscr{D}$ consisting of patterns $\boldsymbol{x}_i$ and labels $y_i = \pm 1$.
- y_i is generated by an unknown target function: $F(\boldsymbol{x}_i) = y_i$.
- The expected off-training set classification error for the k-th learning algorithm is:

$$E_k\{e|F,n\} = \sum_{\boldsymbol{x}\notin\mathscr{D}} p(\boldsymbol{x})\,[1-\delta(F(\boldsymbol{x}),h(\boldsymbol{x}))]\,p_k(h(\boldsymbol{x})|\mathscr{D})$$

where e is the error and $h(\boldsymbol{x})$ the hypothesis on the data.

Off-Training Set Error (cont.)

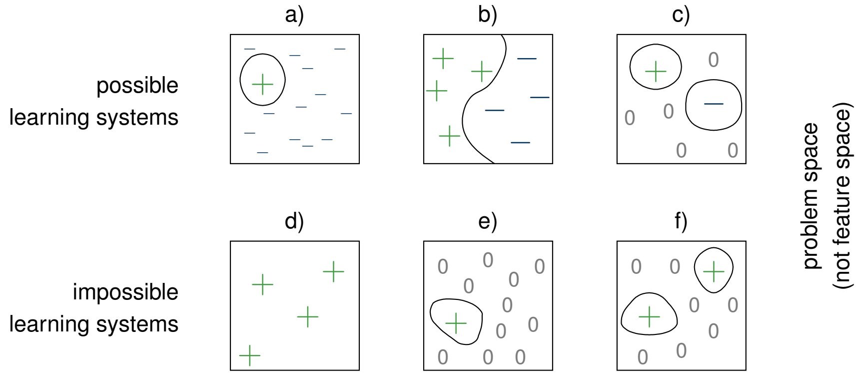


Fig.: Each square represents all possible classification problems. +/— indicates better/worse generalization than the average (adapted from Duda, Hart).

Bias and Variance

- The *No Free Lunch Theorem* states that there is no general best classifier.
- But we have to assess the quality of a learning algorithm in terms of the alignment to the classification problem.
- This can be achieved using the bias-variance relation.

Bias:

- The bias measures the accuracy or quality of the match:
  high bias means poor match.

Variance:

- The variance measures the precision of specificity for the match:
  high variance implies a weak match.

Bias and Variance for Regression

The bias-variance relation is very demonstrative in the context of regression:

- Let $g(\boldsymbol{x};\mathscr{D})$ be the regression function.
- The mean-square deviation from the true function $F(\boldsymbol{x})$ is:

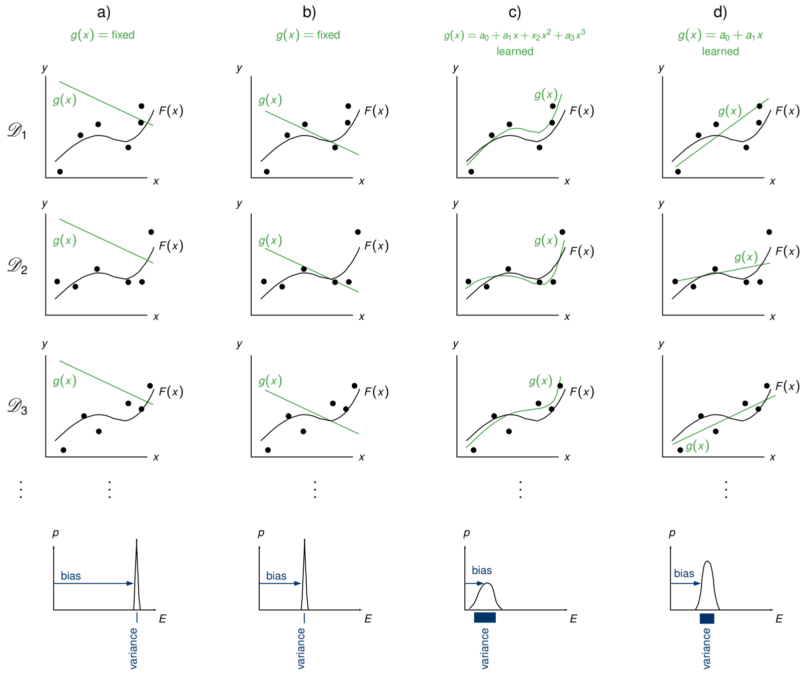
$$
\begin{aligned}
&E_{\mathscr{D}}\left\{\left(g(\boldsymbol{x};\mathscr{D})-F(\boldsymbol{x})\right)^2\right\}\\
&\quad=\underbrace{E_{\mathscr{D}}\left\{g(\boldsymbol{x};\mathscr{D})-F(\boldsymbol{x})\right\}^2}_{(\text{bias})^2}+\underbrace{E_{\mathscr{D}}\left\{\left(g(\boldsymbol{x};\mathscr{D})-E_{\mathscr{D}}\left\{g(\boldsymbol{x};\mathscr{D})\right\}\right)^2\right\}}_{\text{variance}}
\end{aligned}
$$

Bias and Variance for Regression (cont.)

Bias-Variance Trade-Off:

- Methods with high flexibility to adapt to the training data
  - generally have low bias
  - but yield high variance.
- Methods with few parameters and less degrees of freedom
  - tend to have a high bias, as they may not fit the data well.
  - However, this does not change a lot between different data sets, so these methods generally have low variance.
- Unfortunately, we can virtually never get both zero bias and zero variance!
- We need to have as much prior information about the problem as possible to reduce both values.

Bias and Variance for Regression (cont.)



Bias and Variance for Classification

Assuming a two-class classification problem:

- In a two-class problem, the target function changes to:

$$F(\boldsymbol{x}) = p(y = 1|\boldsymbol{x}) = 1 - p(y = -1|\boldsymbol{x})$$

- We cannot compare $g(\boldsymbol{x}; \mathscr{D})$ and $F(\boldsymbol{x})$ based on the mean-square error as in regression.

- For simplicity, let us assume identical priors: $p_1 = p_2 = 0.5$
  - The Bayes discriminant y_B has the threshold 0.5.
  - The Bayes decision boundary is the set of points for which $F(\boldsymbol{x}) = 0.5$.

Bias and Variance for Classification (cont.)

Boundary error

- $p(g(\boldsymbol{x};\mathscr{D}))$ is the pdf of obtaining a particular estimate of the discriminant given $\mathscr{D}$.
- Because of random variations in the training set, the boundary error will depend upon $p(g(\boldsymbol{x};\mathscr{D}))$.

$$p(g(\boldsymbol{x};\mathscr{D}) \neq y_B) = \begin{cases} \displaystyle\int_{0.5}^{\infty} p(g(\boldsymbol{x};\mathscr{D}))\,\mathrm{d}g & \text{if } F(\boldsymbol{x}) < 0.5 \\ \displaystyle\int_{-\infty}^{0.5} p(g(\boldsymbol{x};\mathscr{D}))\,\mathrm{d}g & \text{if } F(\boldsymbol{x}) \leq 0.5 \end{cases}$$

Bias and Variance for Classification (cont.)

- Convenient assumption that $p(g(\boldsymbol{x};\mathscr{D}))$ is a Gaussian:

$$p(g(\boldsymbol{x};\mathscr{D}) \neq y_B) =$$

$$\Phi\left[\underbrace{\operatorname{sgn}\left(F(\boldsymbol{x}) - \frac{1}{2}\right)\cdot\left(E_{\mathscr{D}}\{g(\boldsymbol{x};\mathscr{D})\} - \frac{1}{2}\right)}_{\text{boundary bias}}\cdot\underbrace{\operatorname{var}\Big(\mathrm{g}(\mathrm{x};\mathscr{D})\Big)^{-1/2}}_{\text{variance}}\right]$$

where Φ is a nonlinear function:

$$\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_t^{\infty} e^{-\frac{1}{2}u^2}\,\mathrm{d}u$$

- $p(g(\boldsymbol{x};\mathscr{D}) \neq y_B)$ represents the incorrect estimation of the Bayes boundary.

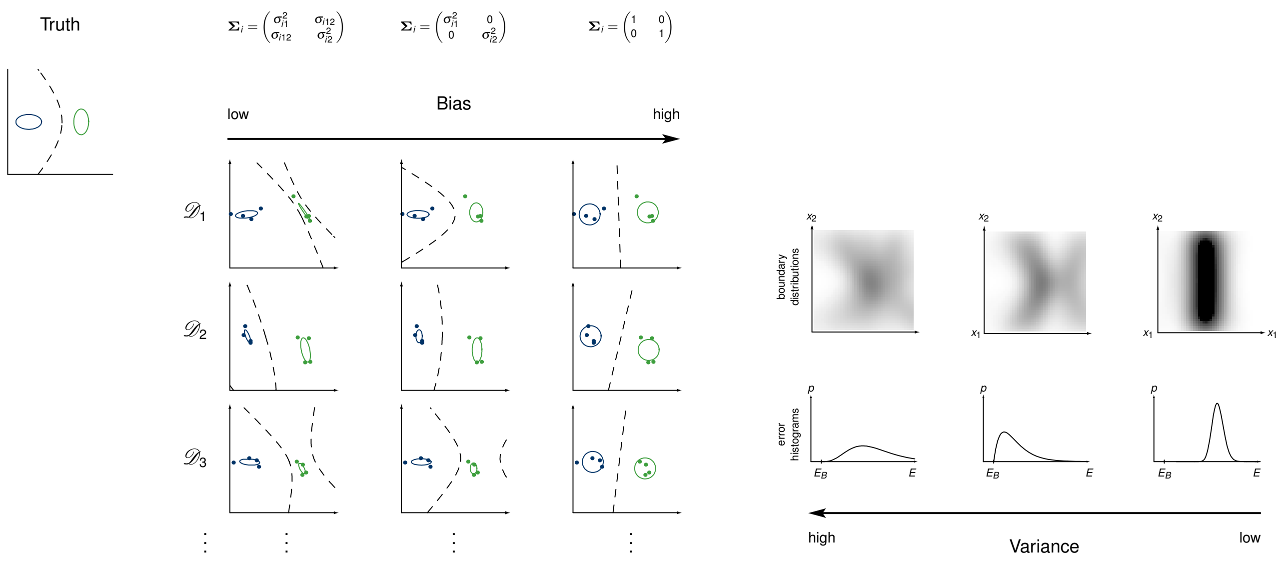
Bias and Variance for Classification (cont.)

Conclusions:

- In regression the bias-variance relation is additive in $(\text{bias})^2$ and variance.
- For classification the relation is multiplicative and nonlinear.
- In classification the sign of the boundary bias affects the role of the variance in the error.
- Therefore, low variance is generally important for accurate classification.

Variance generally dominates bias in classification!

Bias and Variance for Classification (cont.)



Adapted from Duda, Hart

Pattern Recognition Lab

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Next Time in Pattern Recognition

Resampling for Estimating Statistics

Problem:

- Determine the bias and variance for some learning algorithm applied to a new problem with unknown distributions.

From what we have seen so far, bias and variance change with varying samples.

Resampling techniques can be used to yield more informative estimates of a general statistics.

Resampling for Estimating Statistics (cont.)

Formally:

- Suppose we want to estimate a parameter θ that depends on a random sample set $X = (x_1, \dots, x_n)$.
- Assume we have an estimator $\phi_n(X)$ of θ but do not know its distribution.
- Resampling methods try to estimate the bias and variance of $\phi_n(X)$ using subsamples from X.

Jackknife

Let $\text{PS}_i(X)$ be the i-th pseudovalue of $\phi_n(X)$:

$$\begin{aligned}\text{PS}_i(X) &= n\phi_n(X) - (n-1)\phi_{n-1}(X_{(i)}) \\ &= \phi_n(X) - \underbrace{(n-1)(\phi_{n-1}(X_{(i)}) - \phi_n(X))}_{\text{bias}_{\text{jack}}}\end{aligned}$$

where $X_{(i)} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ is the set without the i-th element.

Notes:

- $\text{PS}_i(X)$ can be interpreted as a bias-corrected version of $\phi_n(X)$:
- The bias trend is assumed to be in the estimators from $\phi_{n-1}(X_{(i)})$ to $\phi_n(X)$.

Jackknife (cont.)

Jackknife Principle:

- The pseudovalues $\mathrm{PS}_i(X)$ are treated as independent random variables with mean θ.
- Using the central limit theorem, the ML estimators for the mean μ_{PS} and variance σ^2_{PS} of the pseudovalues are:

$$\mu_{\mathrm{PS}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{PS}_i(X)$$

$$\sigma^2_{\mathrm{PS}} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathrm{PS}_i(X) - \mu_{\mathrm{PS}})^2$$

Jackknife (cont.)

Example

Estimator for the sample mean: $\phi_n(X) = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{X}$

Pseudovalues of $\phi_n(X)$:

$$\mathrm{PS}_i(X) = n\overline{X} - (n-1)\overline{X_{(i)}} = x_i$$

Jackknife estimates:

$$\mu_{\mathrm{PS}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{PS}_i(X) = \overline{X}$$

$$\sigma^2_{\mathrm{PS}} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{X})^2$$

Jackknife (cont.)

Example

Estimator for sample variance: $\phi_n(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2$

Pseudovalues of $\phi_n(X)$:

$$\mathrm{PS}_i(X) = \frac{n}{n-1}(x_i - \overline{X})^2$$

Which implies that:

$$\mu_{\mathrm{PS}} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{PS}_i(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

Interestingly:

- $E\{\phi_n(X)\} = \frac{n-1}{n}\sigma^2$ whereas $E\{\mu_{\mathrm{PS}}\} = \sigma^2$
- μ_{PS} is a bias-corrected version of $\phi_n(X)$

Bootstrap

Literary Sidenote:
The term bootstrap comes from the story: *The adventures of Baron Münchhausen*.

- A *bootstrap* data set is created by randomly selecting n points from the sample set with replacement.
- In *bootstrap estimation* this selection process is independently repeated B times.
- The B bootstrap data sets are treated as independent sets.

Bootstrap (cont.)

The bootstrap estimate of a statistic θ and its variance are the mean of the B estimates $\hat{\theta}^B$ and its variance:

$$\mu_{\text{BS}} = \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_i^B$$

$$\sigma_{\text{BS}}^2 = \frac{1}{B-1}\sum_{i=1}^{B}\left(\hat{\theta}_i^B - \mu_{\text{BS}}\right)^2$$

The bias is the difference between the bootstrap estimate and the estimator $\phi_n(X)$:

$$\text{bias}_{\text{BS}} = \mu_{\text{BS}} - \phi_n(X)$$

Bootstrap (cont.)

Properties of the bootstrap estimate:

- Bootstrapping does not change the prior of the data (choose with replacement).
- The larger the number B, the more will the bootstrap estimate tend towards the true statistic θ.
- In contrast, the jackknife estimator requires exactly n repetitions:
  - less than n repetitions yield poorer estimates
  - more than n repetitions merely duplicate information already provided

Estimating and Comparing Classifiers

Two reasons why we want to know the generalization rate of a classifier on a given problem:

1. to see if the classifier performs well enough to be useful
2. to compare its performance with a competing design

Cross-Validation

- In cross-validation, the training samples are split into two disjoint parts:
  - The first set is the training set used for the traditional training.
  - The second set is the test set used to estimate the classification error.
  - In a second step, both sets are swapped.
  - By that, the classification error can be estimated on the complete data set.
  - Yet training and test set are always disjoint.
- An m-fold cross-validation splits the data into m disjoint sets of size n/m:
  - 1 set is used as test set.
  - The other $m-1$ sets are used for training.
  - Each set is used once for testing.
- In the extreme case of $m = n$, we have a jackknife estimate of the classification accuracy.

Cross-Validation (cont.)

The classifier is trained until a minimum validation error is reached (good generalization vs. overfitting):

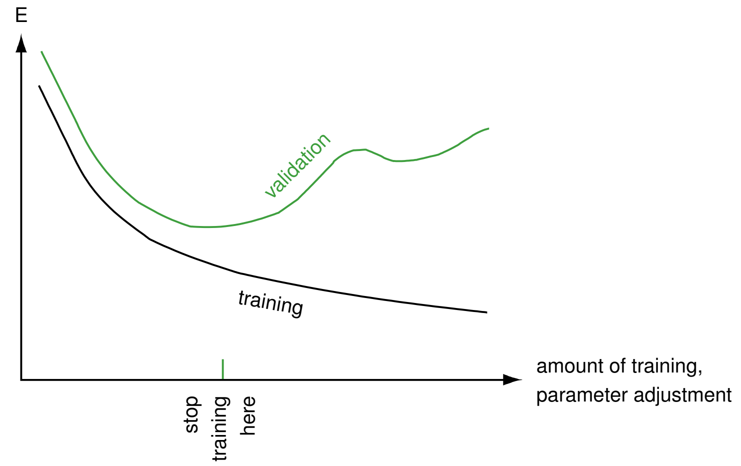


Fig.: The validation error plotted against the amount of training data (adapted from Duda, Hart).

Lessons Learned

- There is no such thing as a free lunch!
- Bias-variance trade-off
- Jackknife
- Bootstrap
- Cross-Validation

Pattern Recognition Lab

FRIEDRICH-ALEXANDER UNIVERSITÄT ERLANGEN-NÜRNBERG
TECHNISCHE FAKULTÄT

Next Time in Pattern Recognition

Further Readings

Examples and various content have been taken from:

- Richard O. Duda, Peter E. Hart, David G. Stork: Pattern Classification, 2nd Edition, John Wiley & Sons, New York, 2000.
- S. Sawyer: Resampling Data: Using a Statistical Jackknife, Washington University, 2005.

Further reading:

- T. Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning, 2nd Edition, Springer, 2009.

Comprehensive Questions

- What is the meaning of the terms bias and variance?
- What is the difference in bias-variance trade-off between regression and classification?
- How do you estimate the bias and variance of a method?
- What is cross-validation and how can it be used to train a classifier?