# Hypothesenbewertungen: Übersicht

Wie kann man Fehler einer Hypothese abschätzen?

Wie kann man einschätzen, ob ein Algorithmus besser ist als ein anderer?

- Trainingsfehler, wirklicher Fehler
- Kreuzvalidierung (Cross Validation)
- Signifikanztests: Vorzeichentest

*Hintergrundinformationen*

- *Schätzer, Konfidenzintervalle*
- *Binomialverteilung, Normalverteilung, Zentraler Grenzwertsatz*
- *Vergleich von Lernverfahren*

# 2 Arten von Fehler

Der **wirkliche Fehler** einer Hypothese $h$ bezüglich einer Zielfunktion $f$ und einer Verteilung $\mathcal{D}$ ist die Wahrscheinlichkeit daß $h$ eine zufällig bezüglich $\mathcal{D}$ gezogene Instanz falsch klassifiziert.

$$\textit{error}_{\mathcal{D}}(h) \equiv \Pr_{x \in \mathcal{D}}[f(x) \neq h(x)]$$

Der **Trainingsfehler** von $h$ bezüglich einer Zielfunktion $f$ und Trainingsmenge $S$ ist der Anteil der Beispiele, die von $h$ falsch klassifiziert werden.

$$\textit{error}_{S}(h) \equiv \frac{1}{|S|} \sum_{x \in S} \delta(f(x) \neq h(x))$$

$\delta(f(x) \neq h(x)) = 1$ falls $f(x) \neq h(x)$ und 0 sonst.

*Wie ist das Verhältnis von error$_S(h)$ und error$_{\mathcal{D}}(h)$?*

# Methoden zur Fehlerabschätzung

Gegeben: Beispielmenge $B$

Naiver Ansatz (nicht anwenden!):

- Wende Lernverfahren auf $B$ an und erzeuge $h$.
- Bestimme $\mathit{error}_B(h)$.

Problem:

- *Bias:* $\mathit{error}_B(h)$ ist zu optimistisch

$$\mathit{bias} \equiv E[\mathit{error}_B(h)] - \mathit{error}_{\mathcal{D}}(h)$$

Für gute ('unbiased') Schätzungen müssen Hypothese und Testmenge unabhängig sein.

# Methoden zur Fehlerabschätzung

Gegeben: Beispielmenge $B$

Besserer Ansatz:

- Teile $B$ auf in Trainingsmenge $T$ und Testmenge $S$
- Wende Lernverfahren auf $T$ an und erzeuge $h$.
- Bestimme *error*$_S(h)$.

Viel besser, aber:

- *Varianz:* Selbst mit unabhängiger Testmenge $S$ kann *error*$_S(h)$ von *error*$_{\mathcal{D}}(h)$ *abweichen*

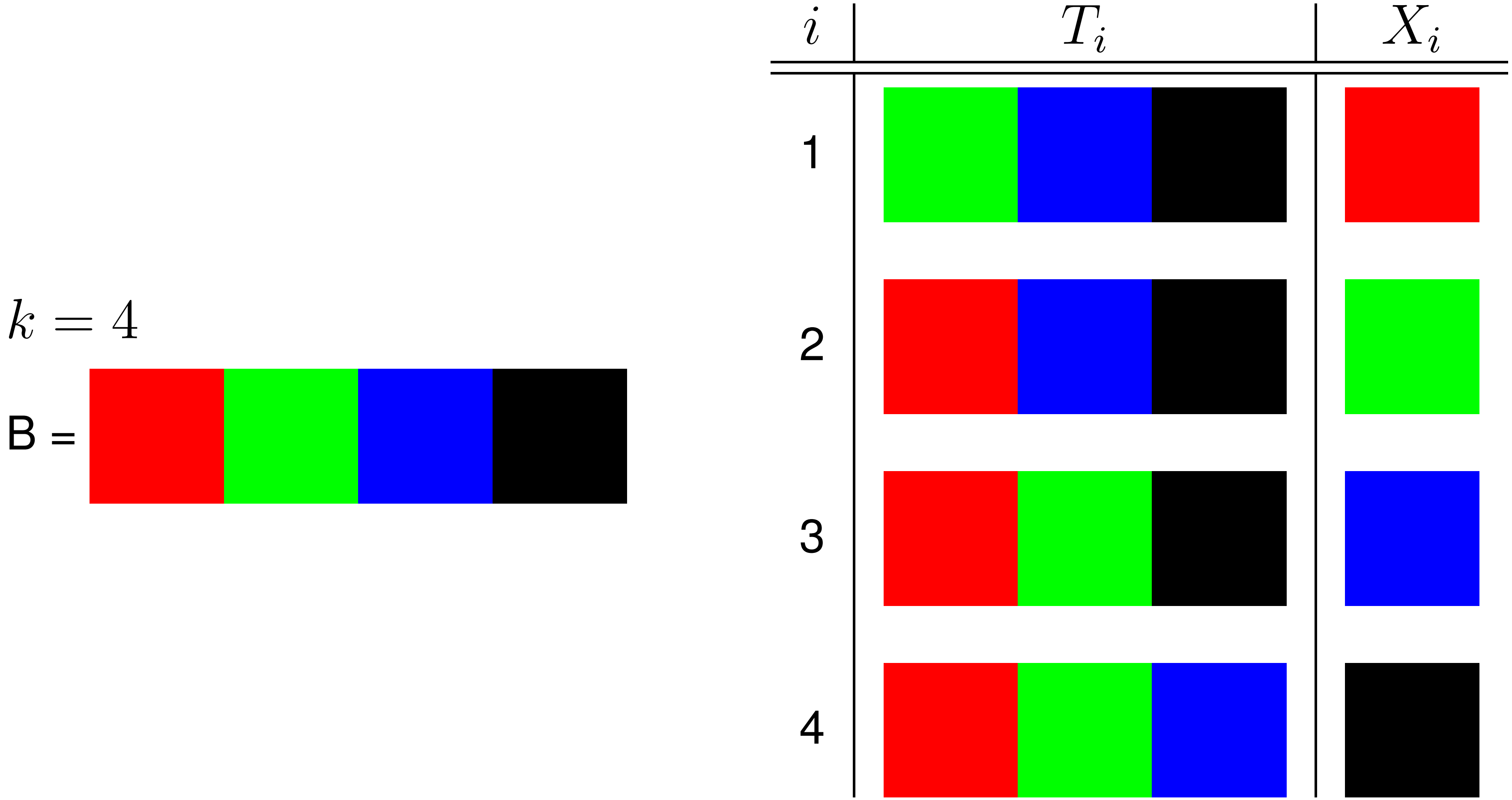
# Methoden zur Fehlerabschätzung: Cross-Validation

Gegeben: Beispielmenge $B$

Cross Validation:

- Teile $B$ auf $k$ disjunkte Mengen $B_1, \ldots, B_k$.
- für $i = 1$ bis $k$: *(Benutze $B_i$ als Testmenge und den Rest als Trainingsmenge)*
  - $T_i = B \setminus B_i, X_i = B_i$
  - Wende Lernverfahren auf $T_i$ an und erzeuge $h_i$.
  - Bestimme $\delta_i = \mathit{error}_{X_i}(h_i)$.
- Wende Lernverfahren auf $B$ an und erzeuge $h$.
- Schätze Fehler von $h$ mit $\frac{1}{k} \cdot \sum_{i=1}^{k} \delta_i$ ab.

# Visualisierung der Cross-Validation



# Anmerkungen zur Cross-Validation

- in Praxis: $k = 10$
  - *tenfold Cross Validation*
- in Praxis: Wiederholung der Prozedur, meist auch wieder 10 Mal
  - *ten tenfold Cross Validation*
- Stratification: Verhältnis der Klassen zueinander in $B$ und in den $B_i$ ist ungefähr gleich
  - Beispiel:
    * $B$ enthält 790 Beispiele der Klasse $a$, 10 Beispiele der Klasse $b$ und 200 Beispiele der Klasse $c$
    * 3-fach Cross Validation
    * Stratification: jedes $B_i$ enthält ungefähr $79\%$ Bsp. der Klasse $a$, $1\%$ Bsp. der Klasse $b$ und $20\%$ Bsp. der Klasse $c$
    * D.h., $B_1$, $B_2$ und $B_3$ enthalten ungefähr $263$ Bsp. der Klasse $a$, $3$ Bsp. der Klasse $b$ und $66$ Bsp. der Klasse $c$
- Wenn $k = |B| \longrightarrow$ *Leave-One-Out*

# Bootstrap

- Idee: ziehe zufällig $n$ Beispiele aus $B$, wobei Beispiele wiederholt gezogen werden können
- Testmenge sind alle diejenigen Beispiele, die nicht für die Traingsmenge gezogen wurden
- Fehlerabschätzung setzt sich zusammen aus Fehlerabschätzung auf Testmenge und auf Trainingsmenge

1. $T \leftarrow$ ziehe zufällig $n$ Beispiele (mit Wiederholung) aus $B$
2. $X \leftarrow B - T$
3. Wende Lernverfahren auf $T$ an und erzeuge $h$.
4. $\mathit{error}(h) \equiv 0,632 \cdot \mathit{error}_X(h) + 0,368 \cdot \mathit{error}_T(h)$

- Woher Faktoren?
  - Wahrscheinlichkeit, daß ein Beispiel in $T$ auftaucht: $\left(1 - \frac{1}{n}\right)^n \approx e^{-1} = 0,368$

# Signifikanztests

Signifikanztest: Verfahren, das anhand der gegebenen Stichprobe entscheidet, ob eine Annahme über eine hypothetische Verteilung mit der konkreten Stichprobe verträglich ist oder ob die hypothetische Verteilung von der wahren Verteilung signifikant (d.h. statistisch gesichert) abweicht.

Gegeben:

- 2 Algorithmen $A$ und $B$
- Bei einer Reihe von Experimenten wurden auf Datenmengen $B_1, \ldots, B_n$ die Fehler Fehler $\delta_1^a, \ldots, \delta_n^a$ der Hypothesen von $A$ und die Fehler $\delta_1^b, \ldots, \delta_n^b$ der Hypothesen von $B$ bestimmt

Frage:
Unterscheiden sich die Verfahren $A$ und $B$ signifikant?

Beispiel:

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\delta_i^a$ | 0,91 | 0,86 | 0,93 | 0,74 | 0,65 | 0,91 | 0,87 | 0,95 | 0,78 | 0,86 | 0,98 | 0,96 | 0,74 | 0,53 | 0,95 | 0,67 | 0,98 | 0,96 | 0,97 | 0,91 |
| $\delta_i^b$ | 0,94 | 0,80 | 0,96 | 0,88 | 0,84 | 0,94 | 0,97 | 0,67 | 0,86 | 0,89 | 0,87 | 0,90 | 0,79 | 0,51 | 0,96 | 0,69 | 0,79 | 0,98 | 0,98 | 0,76 |

# Vorzeichentest

Idee: Unterscheide nur 2 Fälle pro Experiment:

(+) $A$ ist besser als $B$

(-) $B$ ist besser als $A$

---

- Zähle Anzahl der Fälle $+$ $(p_+)$ und der $-$ $(p_-)$
- Bestimme $k$ so, daß $\Pr(p_+ < k$ oder $p_+ > n - k) \leq \alpha$ für ein gegebenes Konfidenzniveau $\alpha$
- Wenn $k \leq p_+ \leq n - k$ gilt, dann ist der Unterschied von $A$ und $B$ nicht signifikant, für $p_+ < k$ und $p_+ > n - k$ ist der Unterschied signifikant.

---

Wie groß muß $k$ sein?

$$\alpha \geq 2 \cdot \sum_{i=0}^{k-1} \binom{n}{i} \cdot \frac{1}{2^n}$$

Für ausreichend große $n$ gilt auch: $k = \frac{1}{2}\left(n - u_{1-\alpha/2} \cdot \sqrt{n}\right)$

| $p$ | $0,80$ | $0,90$ | $0,95$ | $0,975$ | $0,99$ | $0,999$ |
|---|---|---|---|---|---|---|
| $u_p$ | $0,84$ | $1,28$ | $1,64$ | $1,96$ | $2,33$ | $3,09$ |

# Beispiel

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\delta_i^a$ | 0,91 | 0,86 | 0,93 | 0,74 | 0,65 | 0,91 | 0,87 | 0,95 | 0,78 | 0,86 | 0,98 | 0,96 | 0,74 | 0,53 | 0,95 | 0,67 | 0,98 | 0,96 | 0,97 | 0,91 |
| $\delta_i^b$ | 0,94 | 0,80 | 0,96 | 0,88 | 0,84 | 0,94 | 0,97 | 0,67 | 0,86 | 0,89 | 0,87 | 0,90 | 0,79 | 0,51 | 0,96 | 0,69 | 0,79 | 0,98 | 0,98 | 0,76 |
| | - | + | - | - | - | - | - | + | - | - | + | - | - | + | - | - | + | - | - | + |

$$p_+ = 6, p_- = 14$$

Wahrscheinlichkeit der Aussage mind. 95%, d.h. $\alpha = 0,05$

Für $k = 6$: $$2 \cdot \left[\binom{20}{0} + \cdots + \binom{20}{5}\right] \cdot \frac{1}{2^{20}} = 0,041$$

Für $k = 7$: $$2 \cdot \left[\binom{20}{0} + \cdots + \binom{20}{6}\right] \cdot \frac{1}{2^{20}} = 0,115$$

$$\rightarrow k = 6$$

signifikant für $p_+ \in \{0, 1, 2, 3, 4, 5, 15, 16, 17, 18, 19, 20\}$

$A$ und $B$ sind nicht signifikant verschieden.

# Tabelle für $k$ in Abhängigkeit der Stichprobengröße

$N$: Stichprobengröße
$k =$ Wert in entspr. Spalte $+1$

*Vorzeichentest:* Kritische Häufigkeiten $i$ bzw. $N - i$ (s. S. 167)

| $N$ | Irrtumswahrscheinlichkeit 1% | Irrtumswahrscheinlichkeit 5% | $N$ | Irrtumswahrscheinlichkeit 1% | Irrtumswahrscheinlichkeit 5% |
|---|---|---|---|---|---|
| 6 | — | 0 | 41 | 11 | 13 |
| 7 | — | 0 | 42 | 12 | 14 |
| 8 | 0 | 0 | 43 | 12 | 14 |
| 9 | 0 | 1 | 44 | 13 | 15 |
| 10 | 0 | 1 | 45 | 13 | 15 |
| 11 | 0 | 1 | 46 | 13 | 15 |
| 12 | 1 | 2 | 47 | 14 | 16 |
| 13 | 1 | 2 | 48 | 14 | 16 |
| 14 | 1 | 2 | 49 | 15 | 17 |
| 15 | 2 | 3 | 50 | 15 | 17 |
| 16 | 2 | 3 | 51 | 15 | 18 |
| 17 | 2 | 4 | 52 | 16 | 18 |
| 18 | 3 | 4 | 53 | 16 | 18 |
| 19 | 3 | 4 | 54 | 17 | 19 |
| 20 | 3 | 5 | 55 | 17 | 19 |
| 21 | 4 | 5 | 56 | 17 | 20 |
| 22 | 4 | 5 | 57 | 18 | 20 |
| 23 | 4 | 6 | 58 | 18 | 21 |
| 24 | 5 | 6 | 59 | 19 | 21 |
| 25 | 5 | 7 | 60 | 19 | 21 |
| 26 | 6 | 7 | 61 | 20 | 22 |
| 27 | 6 | 7 | 62 | 20 | 22 |
| 28 | 6 | 8 | 63 | 20 | 23 |
| 29 | 7 | 8 | 64 | 21 | 23 |
| 30 | 7 | 9 | 65 | 21 | 24 |
| 31 | 7 | 9 | 66 | 22 | 24 |
| 32 | 8 | 9 | 67 | 22 | 25 |
| 33 | 8 | 10 | 68 | 22 | 25 |
| 34 | 9 | 10 | 69 | 23 | 25 |
| 35 | 9 | 11 | 70 | 23 | 26 |
| 36 | 9 | 11 | 71 | 24 | 26 |
| 37 | 10 | 12 | 72 | 24 | 27 |
| 38 | 10 | 12 | 73 | 25 | 27 |
| 39 | 11 | 12 | 74 | 25 | 28 |
| 40 | 11 | 13 | 75 | 25 | 28 |

# Hintergrundinformation

Ab hier nicht mehr Stoff der Vorlesung

# Was steckt eigentlich dahinter? Ein wenig Statistik

Beispiel:

Hypothese $h$ klassifiziert 12 von 40 Beispielen aus $S$ falsch

$$error_S(h) = \frac{12}{40} = .30$$

Wie groß ist *error*$_{\mathcal{D}}(h)$?

# Schätzer

Experiment:

1. Erzeuge Beispielmenge $S$ der Größe $n$ bezüglich Verteilung $\mathcal{D}$
2. Miß $\mathit{error}_S(h)$

$\mathit{error}_S(h)$ ist Zufallsvariable (d.h. Ergebnis eines Experiments)

$\mathit{error}_S(h)$ ist ein *Schätzer* für $\mathit{error}_{\mathcal{D}}(h)$

Für einen gegebenen $\mathit{error}_S(h)$, was kann man über $\mathit{error}_{\mathcal{D}}(h)$ sagen?

# Konfidenzintervalle

Wenn

- $S$ $n$ unabhängig von $h$ und unabhängig voneinander gezogene Beispiele enthält und
- $n \geq 30$

dann

- liegt $\textit{error}_{\mathcal{D}}(h)$ mit ungefähr 95% Wahrscheinlichkeit im Intervall

$$\textit{error}_S(h) \pm 1.96\sqrt{\frac{\textit{error}_S(h)(1-\textit{error}_S(h))}{n}}$$

# Konfidenzintervalle

Wenn

- $S$ $n$ unabhängig von $h$ und unabhängig voneinander gezogene Beispiele enthält und
- $n \geq 30$

dann

- liegt $error_{\mathcal{D}}(h)$ mit ungefähr $N\%$ Wahrscheinlichkeit im Intervall

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

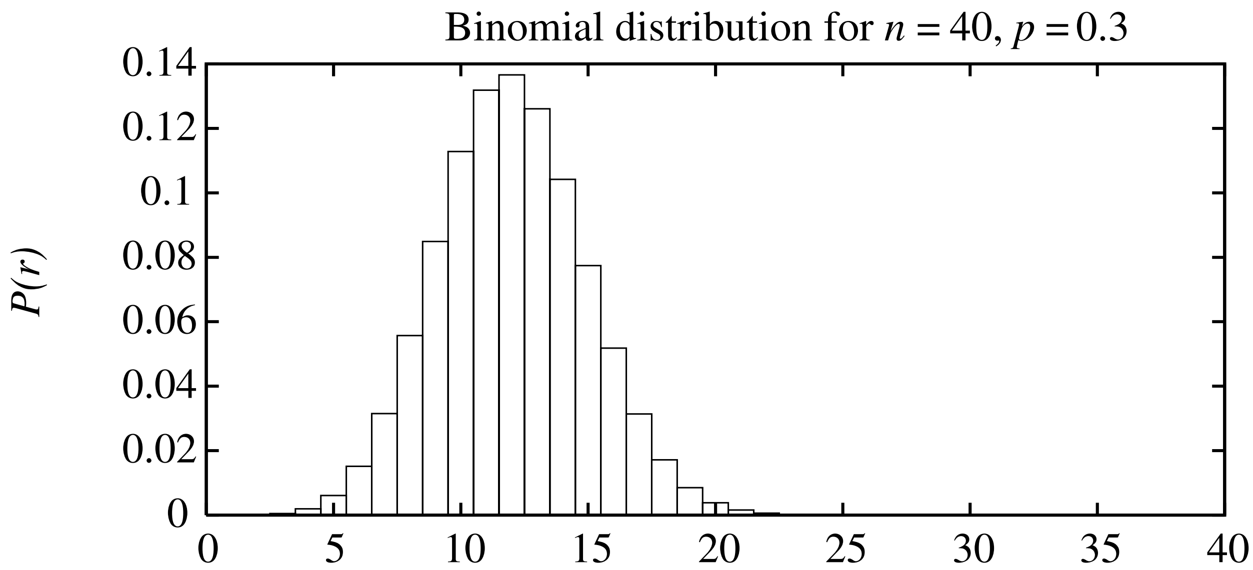
wobei

| $N\%$: | 50% | 68% | 80% | 90% | 95% | 98% | 99% |
|---|---|---|---|---|---|---|---|
| $z_N$: | 0.67 | 1.00 | 1.28 | 1.64 | 1.96 | 2.33 | 2.58 |

# $error_S(h)$ ist eine Zufallsvariable

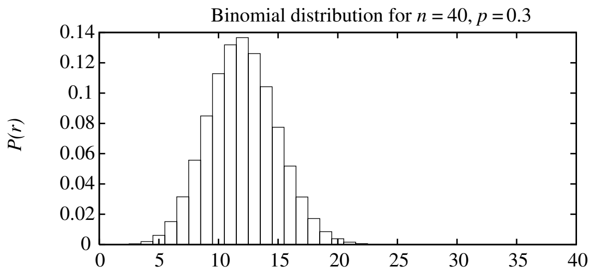
Wiederhole Experiment mit verschiedenen, zufällig erzeugten $S$ der Größe $n$

WK der Beobachtung, daß $r$ Beispiele falsch klassifiziert werden:




$$P(r) = \frac{n!}{r!(n-r)!}\, error_{\mathcal{D}}(h)^r (1 - error_{\mathcal{D}}(h))^{n-r}$$

# Binomialverteilung




$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

Wahrscheinlichkeit $P(r)$ daß $r$ mal ‘Kopf’ bei $n$ Münzwürfen auftritt, falls $p = \Pr(\textit{Kopf})$

- Erwartungswert $E[X]$ von $X$ ist $\boxed{E[X] \equiv \sum_{i=0}^{n} iP(i) = np}$
- Varianz von $X$ ist $\boxed{Var(X) \equiv E[(X - E[X])^2] = np(1-p)}$
- Standardabweichung $\sigma_X$ von $X$ ist $\boxed{\sigma_X \equiv \sqrt{E[(X - E[X])^2]} = \sqrt{np(1-p)}}$

# Normalverteilung approximiert Binomialverteilung

$\textit{error}_S(h)$ folgt einer *Binomialverteilung*, wobei

- Erwartungswert $E[\textit{error}_S(h)] = \textit{error}_{\mathcal{D}}(h)$
- Standardabweichung $\sigma_{\textit{error}_S(h)}$

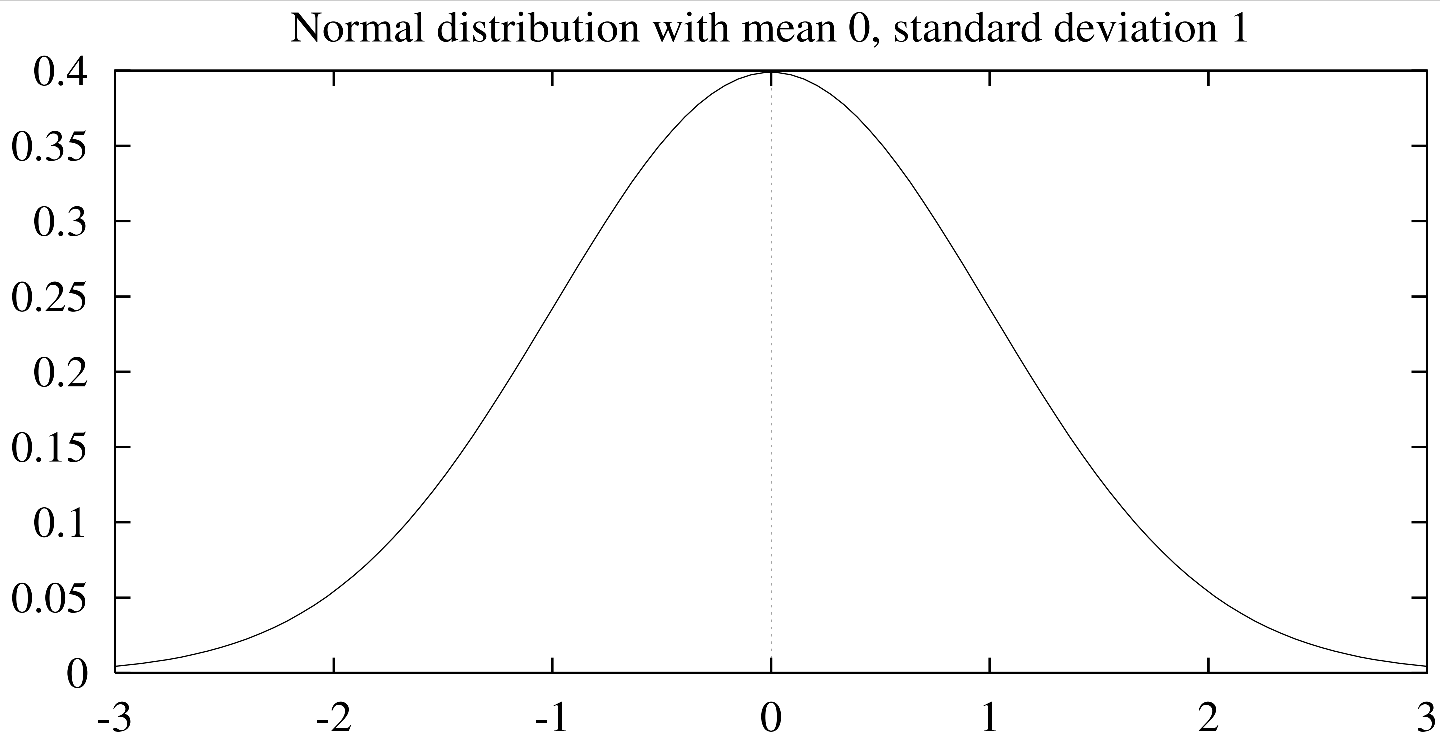
$$\sigma_{\textit{error}_S(h)} = \sqrt{\frac{\textit{error}_{\mathcal{D}}(h)(1 - \textit{error}_{\mathcal{D}}(h))}{n}}$$

Dies kann durch eine *Normalverteilung* approximiert werden, wobei

- Erwartungswert $E[\textit{error}_S(h)] = \textit{error}_{\mathcal{D}}(h)$
- Standardabweichung $\sigma_{\textit{error}_S(h)}$

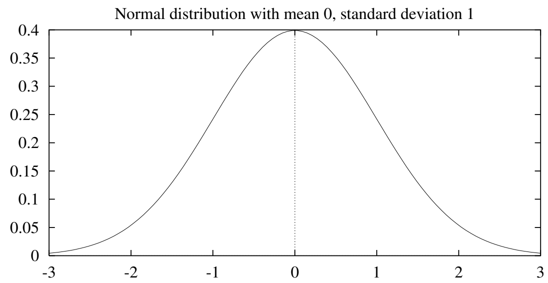
$$\sigma_{\textit{error}_S(h)} \approx \sqrt{\frac{\textit{error}_S(h)(1 - \textit{error}_S(h))}{n}}$$

# Normalverteilung




$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

# Normalverteilung



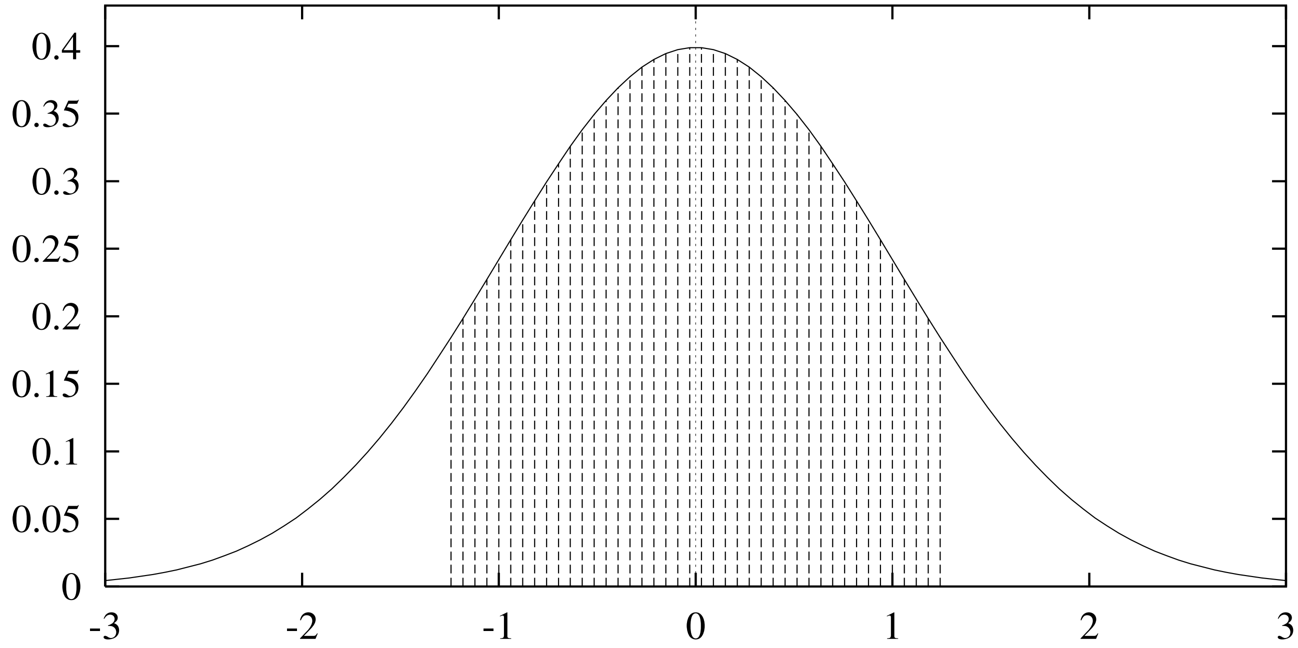

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Wahrscheinlichkeit, daß $X$ in das Intervall $(a, b)$ fällt ist gegeben durch

$$\int_a^b p(x)dx$$

- Erwartungswert von $X$, $E[X]$, ist $\boxed{E[X] = \mu}$
- Varianz von $X$ ist $\boxed{Var(X) = \sigma^2}$
- Standardabweichung von $X$, $\sigma_X$, ist $\boxed{\sigma_X = \sigma}$

# Normalverteilung




80% des Flächeninhalts (Wahrscheinlichkeit) liegt im Bereich $\mu \pm 1.28\sigma$

N% des Flächeninhalts (Wahrscheinlichkeit) liegt im Bereich $\mu \pm z_N\sigma$ mit

| $N\%$: | 50% | 68% | 80% | 90% | 95% | 98% | 99% |
|---|---|---|---|---|---|---|---|
| $z_N$: | 0.67 | 1.00 | 1.28 | 1.64 | 1.96 | 2.33 | 2.58 |

# Zentraler Grenzwertsatz

Betrachte Menge von unabhängigen, gleich verteilten Zufallsvariablen $Y_1 \ldots Y_n$, die alle von der gleichen, beliebigen Verteilung mit Erwartungswert $\mu$ und endlicher Varianz $\sigma^2$ erzeugt werden. Definiere Erwartungswert

$$\bar{Y} \equiv \frac{1}{n} \sum_{i=1}^{n} Y_i$$

**Zentraler Grenzwertsatz.** Für $n \longrightarrow \infty$ nähert sich die Verteilung $\bar{Y}$ einer Normalverteilung mit Erwartungswert $\mu$ und Varianz $\frac{\sigma^2}{n}$ an.

# Berechnung von Konfidenzintervallen

1. Wähle abzuschätzenden Parameter $p$
   - *error*$_{\mathcal{D}}(h)$
2. Wähle Schätzer
   - *error*$_S(h)$
3. Bestimme Wahrscheinlichkeitsverteilung des Schätzers
   - *error*$_S(h)$ binomialverteilt, nähert Normalverteilung an falls $n \geq 30$
4. Finde Intervall $(L, U)$ so daß N% der Wahrscheinlichkeitsmasse in das Intervall fällt
   - Benutze Tabelle für $z_N$

# Differenz zwischen Hypothesen

Teste $h_1$ auf Menge $S_1$, teste $h_2$ auf $S_2$

1. Wähle abzuschätzenden Parameter

$$d \equiv \mathit{error}_{\mathcal{D}}(h_1) - \mathit{error}_{\mathcal{D}}(h_2)$$

2. Wähle Schätzer

$$\hat{d} \equiv \mathit{error}_{S_1}(h_1) - \mathit{error}_{S_2}(h_2)$$

3. Bestimme Wahrscheinlichkeitsverteilung des Schätzers

$$\sigma_{\hat{d}} \approx \sqrt{\frac{\mathit{error}_{S_1}(h_1)(1-\mathit{error}_{S_1}(h_1))}{n_1} + \frac{\mathit{error}_{S_2}(h_2)(1-\mathit{error}_{S_2}(h_2))}{n_2}}$$

4. Finde Intervall $(L, U)$ so daß N% der Wahrscheinlichkeitsmasse hineinfällt

$$\hat{d} \pm z_N \sqrt{\frac{\mathit{error}_{S_1}(h_1)(1-\mathit{error}_{S_1}(h_1))}{n_1} + \frac{\mathit{error}_{S_2}(h_2)(1-\mathit{error}_{S_2}(h_2))}{n_2}}$$

# Vergleich von Lernalgorithmen $A$ und $B$

Wir wollen folgendes abschätzen:

$$E_{S \subset \mathcal{D}}[\textit{error}_{\mathcal{D}}(A(S)) - \textit{error}_{\mathcal{D}}(B(S))]$$

*($L(S)$ ist die Hypothese des Algorithmus $L$ auf den Trainingsdaten $S$)*

D.h. abzuschätzen ist die erwartete Differenz zwischen den wirklichen Fehlern von $A$ und $B$, wenn die Trainingsmengen $S$ zufällig bezüglich der Verteilung $\mathcal{D}$ gezogen werden.

Beschränkte Datenmenge $B$: Was ist ein guter Schätzer?

- Teile $B$ in Trainingsmenge $T$ und Testmenge $X$ auf und bestimme

$$\textit{error}_X(A(T)) - \textit{error}_X(B(T))$$

- noch besser: wiederhole dies immer wieder und bilde Durchschnitte

# Vergleich von Lernalgorithmen: Cross-Validation

Gegeben: Beispielmenge $B$, 2 Lernverfahren $A$ und $B$

Cross Validation:

- Teile $B$ auf $k$ disjunkte Mengen $B_1, \ldots, B_k$.
- für $i = 1$ bis $k$: *(Benutze $T_i$ als Testmenge und den Rest als Trainingsmenge)*
  - $T_i = B \setminus B_i$, $X_i = B_i$
  - Wende Lernverfahren $A$ und $B$ auf $T_i$ an und erzeuge $h_i^a$ und $h_i^b$.
  - Bestimme $\delta_i = \mathit{error}_{X_i}(h_i^a) - \mathit{error}_{X_i}(h_i^b)$.
- Gib $\bar{\delta}$ mit $\bar{\delta} = \frac{1}{k} \cdot \sum_{i=1}^{k} \delta_i$ aus.

# Anmerkungen zur Cross-Validation

Abschätzung des $N$% Konfidenzintervalls für $\delta$:

$$\bar{\delta} \pm t_{N,k-1} \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^{k} (\delta_i - \bar{\delta})^2}$$

$t_{N,k-1}$: Faktor analog $z_N$, aber hier für *Student*sche $t$-Verteilung (mit $k-1$ Freiheitsgraden) $\longrightarrow$ Tabellen

# Changelog

Folie 7: Korrekter Name ist Leave-One-Out, Nicht Holdout-Testing

Folie 11: Typo in Formeln korrigiert, Rechenfehler im Beispiel korrigiert (Faktor 2 wurde vergessen)

Folie 13: hinzugefügt

Folie 5: Es muß heißen: *Benutze* $B_i$ *als Testmenge* und nicht $T_i$