# Maschinelles Lernen zur Vorhersage proteotypischer Peptide

## Bachelorarbeit Bioinformatik

Till Helge Helwig

**Eberhard-Karls-Universität Tübingen**
Fakultät für Informations- und Kognitionswissenschaft
Wilhelm-Schickard-Institut für Informatik

20. August 2008

# Überblick

1. Motivation
2. Grundlagen
3. Datensätze & Methoden
4. Implementierung
5. Ergebnisse
6. Auswertung & Ausblick

# Überblick

1 Motivation

2 Grundlagen

3 Datensätze & Methoden

4 Implementierung

5 Ergebnisse

6 Auswertung & Ausblick

# Motivation

- **Protein-Identifikation** spielt eine große Rolle in vielen verschiedenen Disziplinen, z.B.:
  - Wirkstoffentwicklung
  - Krebsforschung
- Bisherige Verfahren mittels Massenspektrometrie sind sehr aufwändig.
- **Konstengünstigere und schnellere Verfahren** sind von großem Interesse.

# Überblick

1 Motivation

2 Grundlagen

3 Datensätze & Methoden

4 Implementierung

5 Ergebnisse

6 Auswertung & Ausblick

# Proteotypische Peptide

## Grundlegende Überlegung

### Generelle Annahme

Jedes Peptid aus einer Probe wird bei einem massenspektrometrischen Experiment mit **gleicher** Wahrscheinlichkeit gemessen.

# Proteotypische Peptide
## Grundlegende Überlegung

### Generelle Annahme

Jedes Peptid aus einer Probe wird bei einem massenspektrometrischen Experiment mit **gleicher** Wahrscheinlichkeit gemessen.

### Neuere Forschungsergebnisse

Einzelne Peptide einer Probe werden mit **größerer** Wahrscheinlichkeit identifiziert, als die übrigen. Man nennt sie ***proteotypisch***.

# Proteotypische Peptide
## Bedeutung und Folgen

- Protein-Identifikation anhand **eindeutiger Peptide** ist einfacher, zuverlässiger und schneller, als bisherige Verfahren.
- Ergebnisse, die bisher als mögliche *false positives* **verworfen** wurden, sind jetzt von **großem Wert**, wenn sie proteotypische Peptide beinhalten.
- Spezielle Datenbanken für diese neue Methode sind nötig, aber häufig noch **nicht vorhanden**.
- Eine **schnelle und zuverlässige Vorhersage** proteotypischer Peptide für eine Proteinsequenz ist wichtig, um das Verfahren universell einsetzen zu können.

# Maschinelles Lernen

## Grundidee

Eine virtuelle Maschine wird darauf trainiert, die **unbekannte Hintergrundverteilung** der Eingabe-Daten so gut wie möglich zu approximieren.

# Maschinelles Lernen

## Grundidee

Eine virtuelle Maschine wird darauf trainiert, die **unbekannte Hintergrundverteilung** der Eingabe-Daten so gut wie möglich zu approximieren.

- Für das Training werden positive und negative Beispiele aus **bekannten Datensätzen** benutzt.

# Maschinelles Lernen

## Grundidee

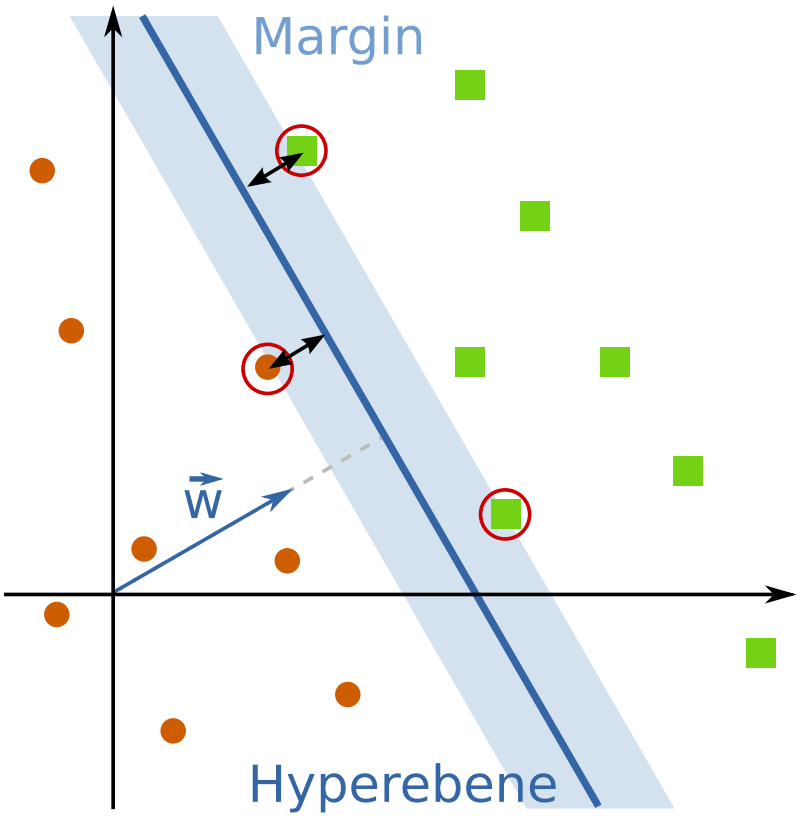
Eine virtuelle Maschine wird darauf trainiert, die **unbekannte Hintergrundverteilung** der Eingabe-Daten so gut wie möglich zu approximieren.

- Für das Training werden positive und negative Beispiele aus **bekannten Datensätzen** benutzt.
- Nach dem Training wird mit weitere Beispielen **getestet**, ob die Maschine diese korrekt klassifiziert.

# Maschinelles Lernen
## Support Vektor Maschinen

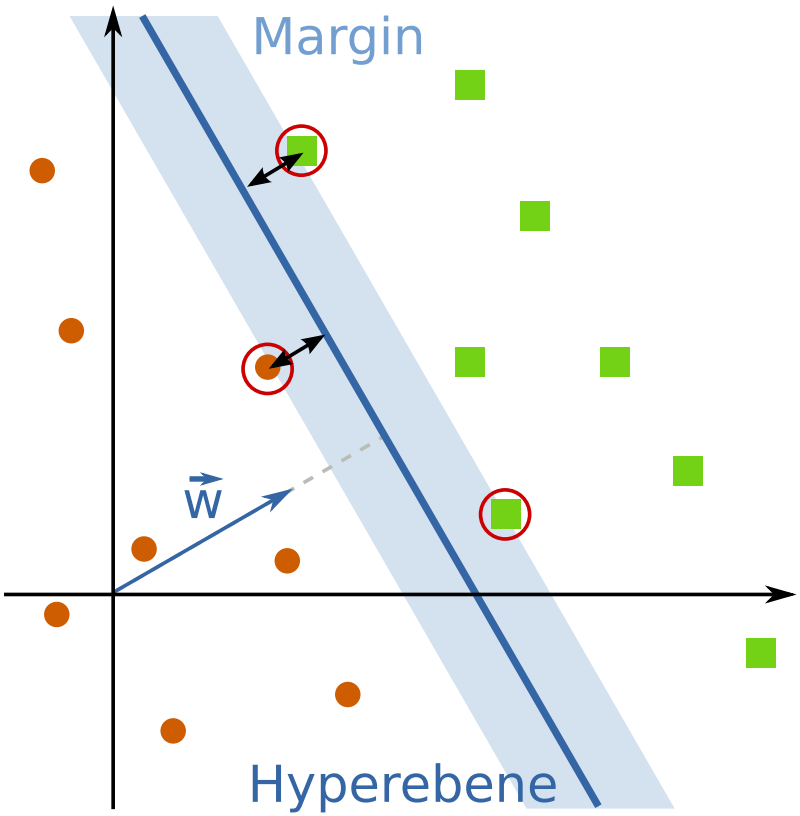
- Die Eingabedaten werden in einen **höherdimensionalen Feature-Raum** abgebildet.

# Maschinelles Lernen
## Support Vektor Maschinen




- Die Eingabedaten werden in einen **höherdimensionalen Feature-Raum** abgebildet.
- Durch Anpassung einer **Hyperbene** werden die Punkte dort in positive und negative Beispiele sortiert.

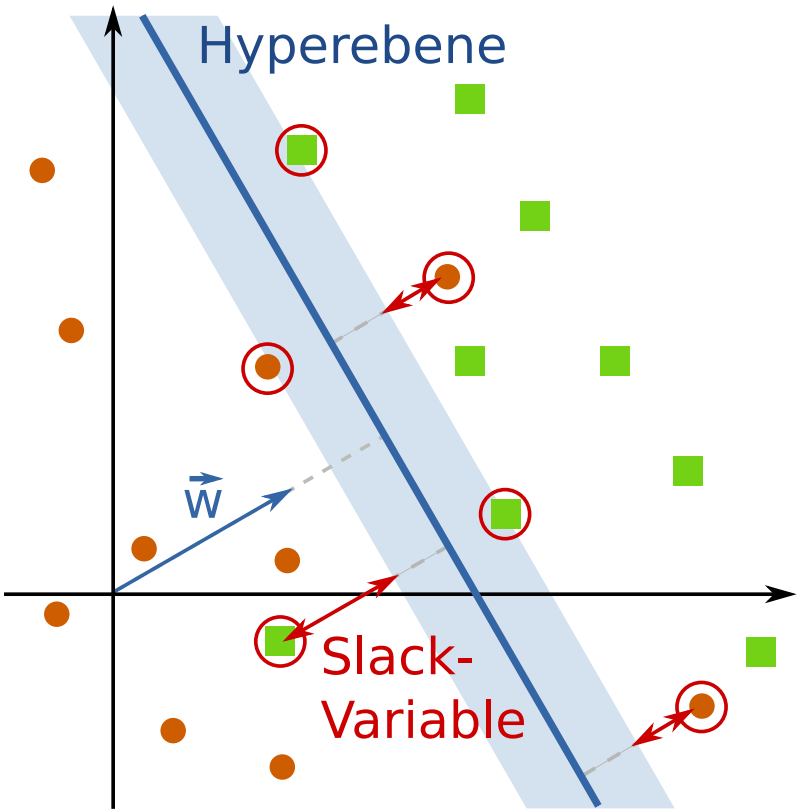
# Maschinelles Lernen
## Support Vektor Maschinen




- Die Eingabedaten werden in einen **höherdimensionalen Feature-Raum** abgebildet.
- Durch Anpassung einer **Hyperbene** werden die Punkte dort in positive und negative Beispiele sortiert.
- Einführung eines **Margins** und dessen Maximierung sorgt für generelle Nutzbarkeit.

# Maschinelles Lernen

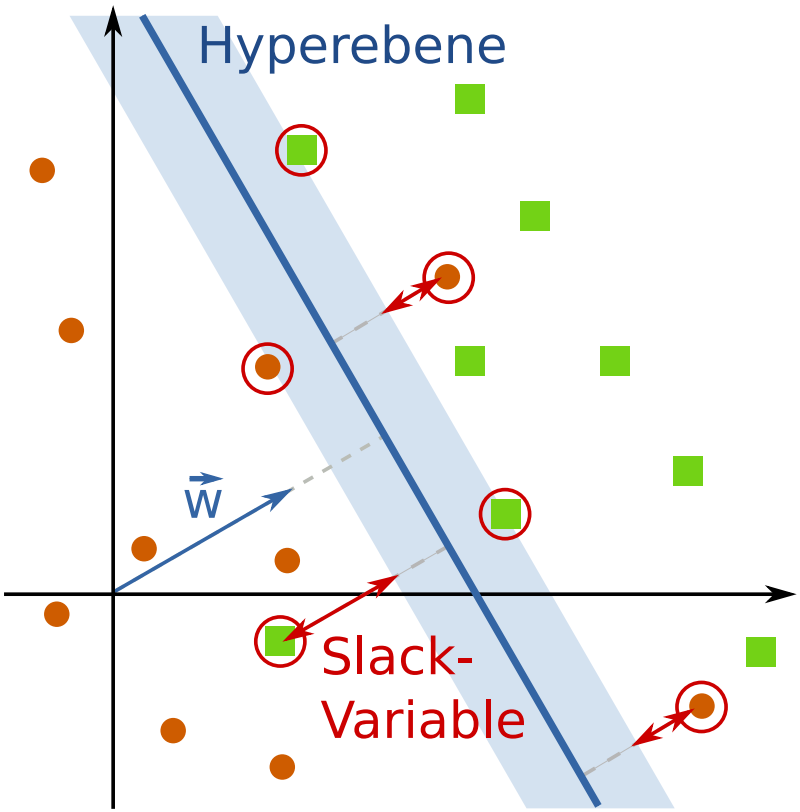
## Support Vektor Maschinen – Slack-Variablen




- Nicht immer ist eine vollständige Trennung der Punkte möglich.
- Mittels **Slack-Variablen** wird erlaubt, dass Punkte falsch klassifiziert werden können.

# Maschinelles Lernen

## Support Vektor Maschinen – Slack-Variablen




- Nicht immer ist eine vollständige Trennung der Punkte möglich.
- Mittels **Slack-Variablen** wird erlaubt, dass Punkte falsch klassifiziert werden können.
- Für die falsche Klassifizierung wird eine **Strafe** eingeführt, um unnötige Slack-Variablen zu verhindern.

# Maschinelles Lernen

## Support Vektor Maschinen – Kernfunktionen und Optimierungsproblem

- Mittels **Kernfunktionen** werden die Daten aus dem Eingabe-Raum in den Feature-Raum abgebildet.
- Die lineare Entscheidungsfunktion des trainierten Klassifikators entspricht somit einer **komplexen Funktion** im Eingabe-Raum.
- Es ergibt sich ein **Optimierungsproblem**, dessen Parameter durch die SVM möglichst günstig bestimmt werden, ohne die Nebenbedingungen zu verletzen.

# Überblick

1 Motivation

2 Grundlagen

3 Datensätze & Methoden

4 Implementierung

5 Ergebnisse

6 Auswertung & Ausblick

# Verfahren und Daten von Mallick et Al.

# Verfahren und Daten von Tang et Al.

# Verfahren von Lu et Al.

# Überblick

1 Motivation

2 Grundlagen

3 Datensätze & Methoden

4 Implementierung

5 Ergebnisse

6 Auswertung & Ausblick

# Implementierung

# Überblick

1 Motivation

2 Grundlagen

3 Datensätze & Methoden

4 Implementierung

5 Ergebnisse

6 Auswertung & Ausblick

# Ergebnisse

# Überblick

1 Motivation

2 Grundlagen

3 Datensätze & Methoden

4 Implementierung

5 Ergebnisse

6 Auswertung & Ausblick

# Auswertung & Ausblick