




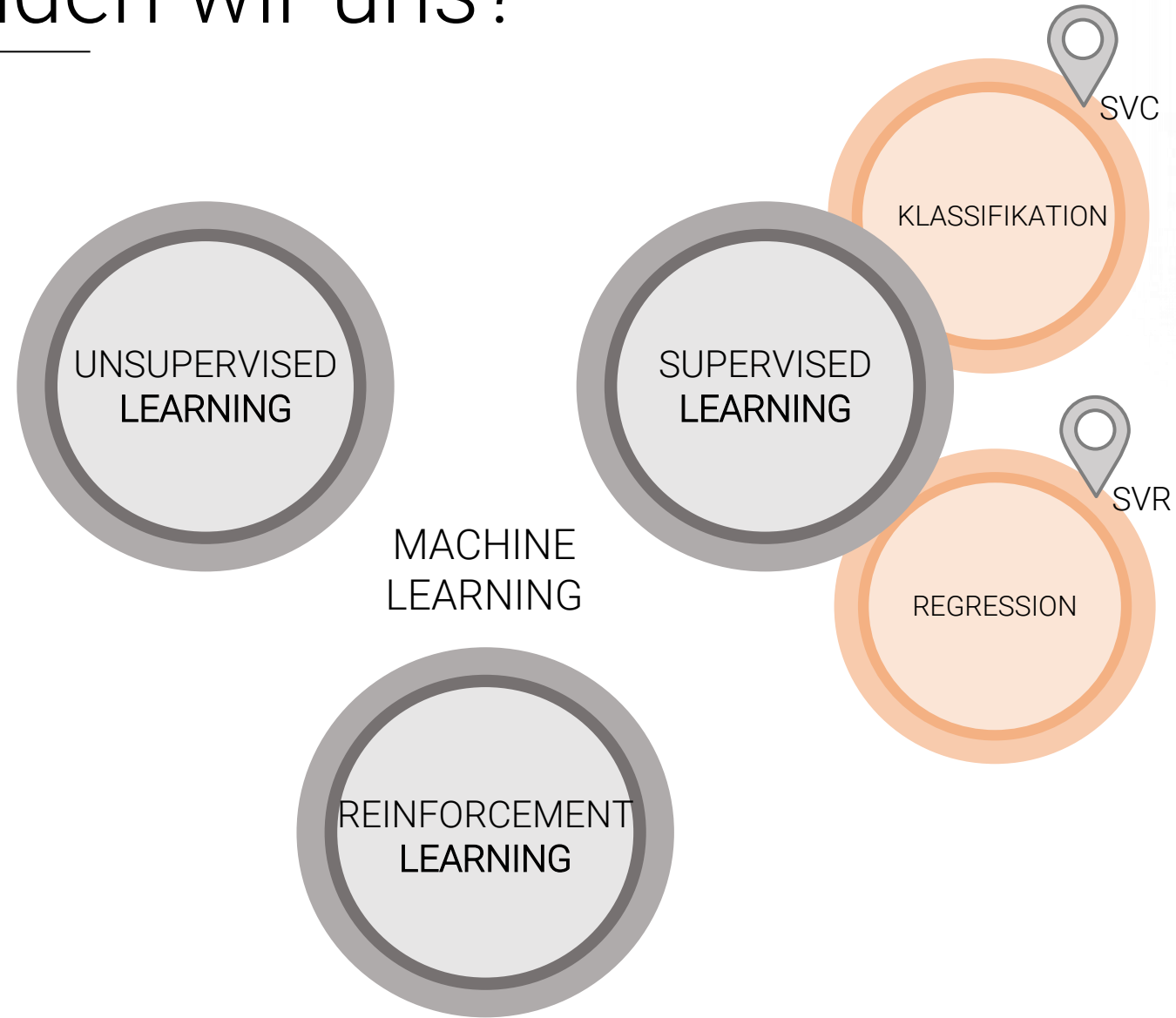
Modeling

Supervised Learning: Support Vector Machines

Agenda

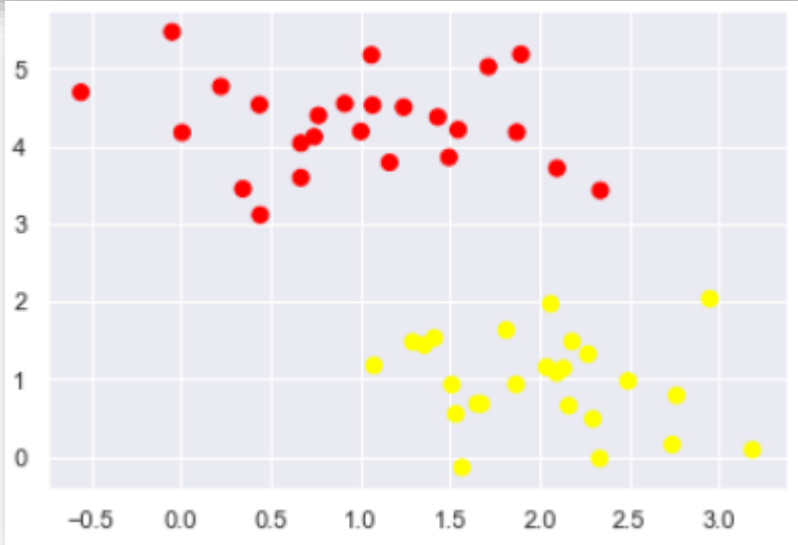
- 
1. SVM: Margin Maximierung und Support Vectors
 2. Mathematischer Hintergrund zu SVMs
 3. Soft-Margin SVM und Optimierung
 4. Nicht-lineare Probleme und Kernel-Trick

SVM: wo befinden wir uns?



SVM: allgemein

```
1 from sklearn.datasets.samples_generator import make_blobs
2 X, y = make_blobs(n_samples=50, centers=2,
3                   random_state=0, cluster_std=0.60)
4 plt.scatter(X[:, 0], X[:, 1], c=y, s=50, cmap='autumn');
```



Unsere Ausgangssituation:

Zwei Klassen in einem 2D-Feature-Raum, die wir mit einem Klassifikator zu trennen versuchen



Was denken Sie?

Welches Problem sehen wir hier sofort?

- Support Vector Machines (**SVM**) sind eine der bekanntesten Modelle im Machine Learning
- Vor allem für **klein- und mittelgroße** Datensätze geeignet
- Sie sind flexibel und können für Klassifikations- und Regressionsprobleme verwendet werden
- SVMs stellen im Kontext der Klassifikation *diskriminative* Klassifikatoren dar



Was denken Sie?

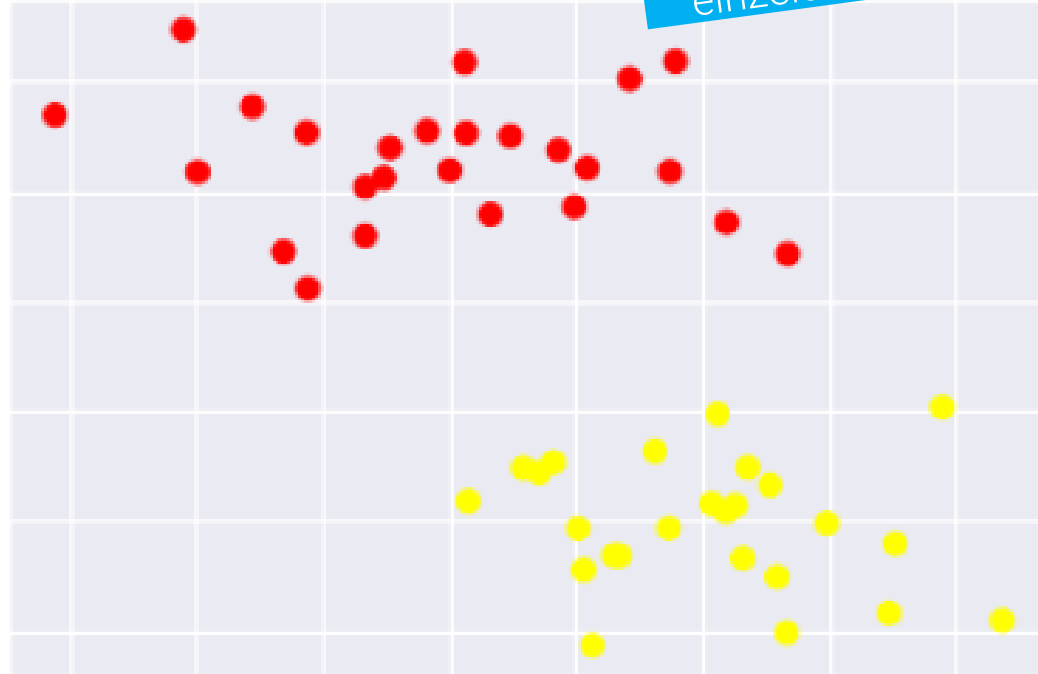
Was heißt diskriminativ?

Wir als Klassifikatoren

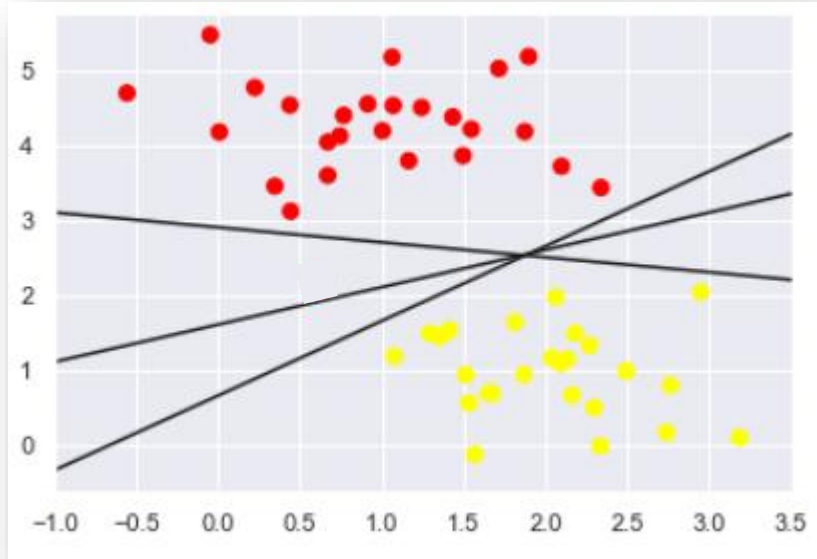


Was denken Sie?

Wie würden Sie – als Klassifikator
– eine optimale Trennlinie
einzeichnen? (Zoom: zeichnen)



SVM: allgemein



- Wir können also beliebig viele Geraden einzeichnen, die unsere Klassen trennen
- Alle drei Geraden separieren unsere Klassen quasi ideal
- Wie sieht es nun mit **neuen Datenpunkten** aus? Je nachdem welche Gerade wir in unsere Ausgangssituation legen, können neue Datenpunkte unterschiedlich klassifiziert werden

```
1 xfit = np.linspace(-1, 3.5)
2 plt.scatter(X[:, 0], X[:, 1], c=y, s=50, cmap='autumn')
3 plt.plot([0.6], [2.1], 'x', color='red', markeredgewidth=2, markersize=10)
4
5 for m, b in [(1, 0.65), (0.5, 1.6), (-0.2, 2.9)]:
6     plt.plot(xfit, m * xfit + b, '-k')
7
8 plt.xlim(-1, 3.5);
```



Was denken Sie?

Wie würden Sie intuitiv das „x“ zuordnen – je nach Gerade?



Was denken Sie?

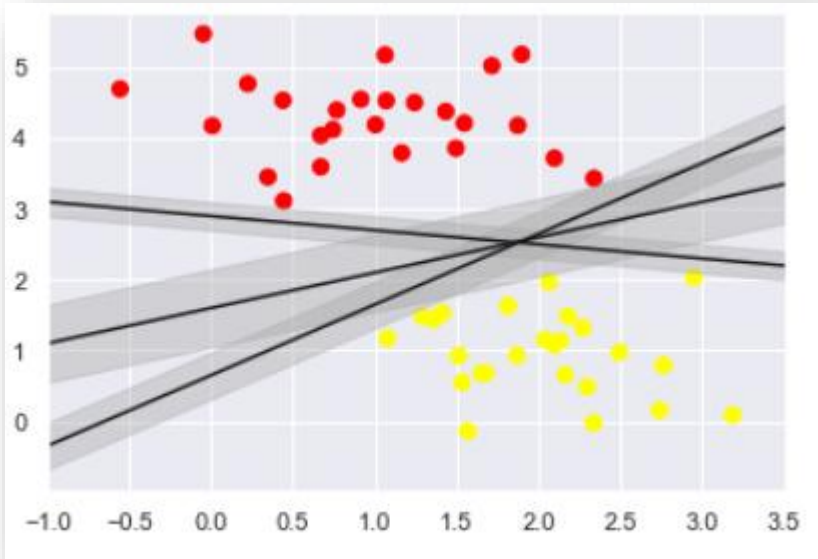
Wie würden Sie intuitiv entscheiden, welche Gerade die beste ist?

SVM: Margin Maximierung



Was denken Sie?

Was könnte man noch tun – außer nur eine Gerade zu zeichnen?



```
1 xfit = np.linspace(-1, 3.5)
2 plt.scatter(X[:, 0], X[:, 1], c=y, s=50, cmap='autumn')
3
4 for m, b, d in [(1, 0.65, 0.33), (0.5, 1.6, 0.55), (-0.2, 2.9, 0.2)]:
5     yfit = m * xfit + b
6     plt.plot(xfit, yfit, '-k')
7     plt.fill_between(xfit, yfit - d, yfit + d, edgecolor='none',
8                     color='AAAAAA', alpha=0.4)
9
10 plt.xlim(-1, 3.5);
```

- Anstatt nur eine bloße Gerade zur Trennung der Klassen zu nutzen, wird ein sog. **Margin** (dt. Rand, Spanne) um die Gerade gelegt
- Dieser Margin wird bis zu den **nächstliegenden** Punkten der Klassen erweitert
- Die **optimale** Gerade ist dann die mit dem **größten** Margin

→ SVMs sind sog. *Maximum Margin Estimators*



Was denken Sie?

Welche Gerade wird wohl die beste sein – und warum?

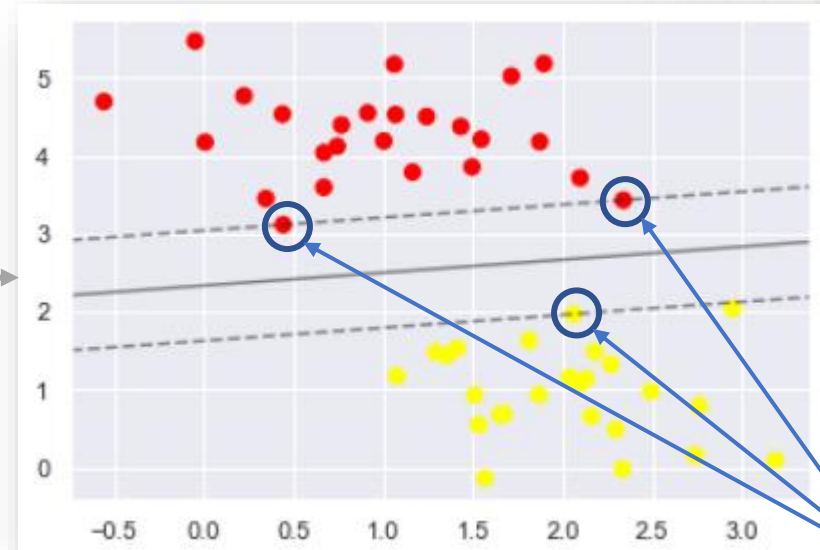
SVM: Fitting/Training

Modell

```
1 from sklearn.svm import SVC # "Support vector classifier"
2 model = SVC(kernel='linear', C=1E10)
3 model.fit(X, y)
```



Was denken Sie?
Können Sie sich vorstellen,
woher das „Support
Vector“ in SVM kommt?



Support Vectors

- Wie wir es bei der Einführung der `sklearn`-Syntax gelernt haben instanziiieren wir das SVM Modell

`SVC(kernel=<kernel_type>, C=<regularization_parameter_value>)`

Was `kernel` und `C` repräsentieren lernen wir später

- Dann können wir die `.fit()`-Methode verwenden, um das Modell auf unsere Daten in `x` (Features) und `y` (Labels) anzupassen/fitten/trainieren
- Dadurch erhalten wir die Trennlinie mit dem **maximalen Margin**, die unsere beiden Klassen am besten trennt
- `SVC` bedeutet **Support Vector Classifier** – SVMs können also auch **regressiv** verwendet werden, sehen wir später

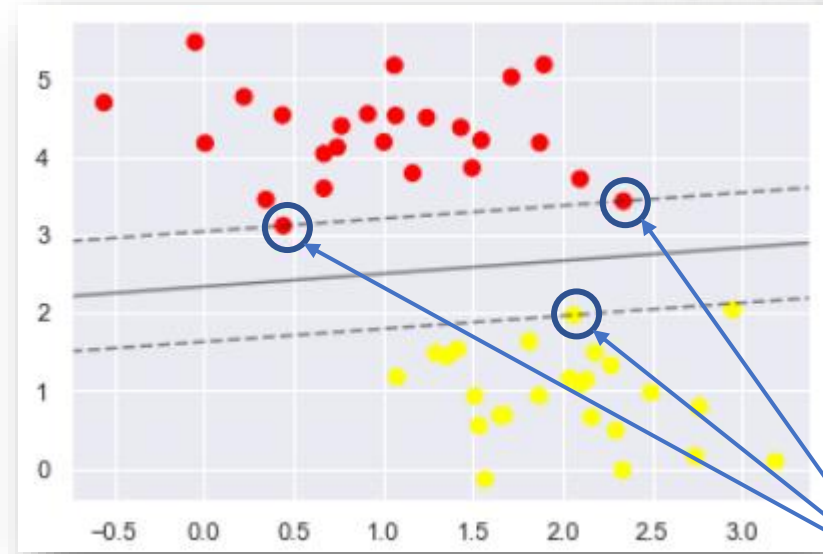
SVM: Support Vectors



Was denken Sie?
Welchen? Von was
hängt das Modell ab?

```
1 model.support_vectors_  
  
array([[ 0.44359863,  3.11530945],  
       [ 2.33812285,  3.43116792],  
       [ 2.06156753,  1.96918596]])
```

- Diese Support Vectors stellen einen entscheidenden Aspekt dieses ML-Modells dar
- Der Fit des Modells hängt **nur** von diesen Datenpunkten ab!
- Wir können über das Attribut `.support_vectors_` auf diese zugreifen



Support Vectors



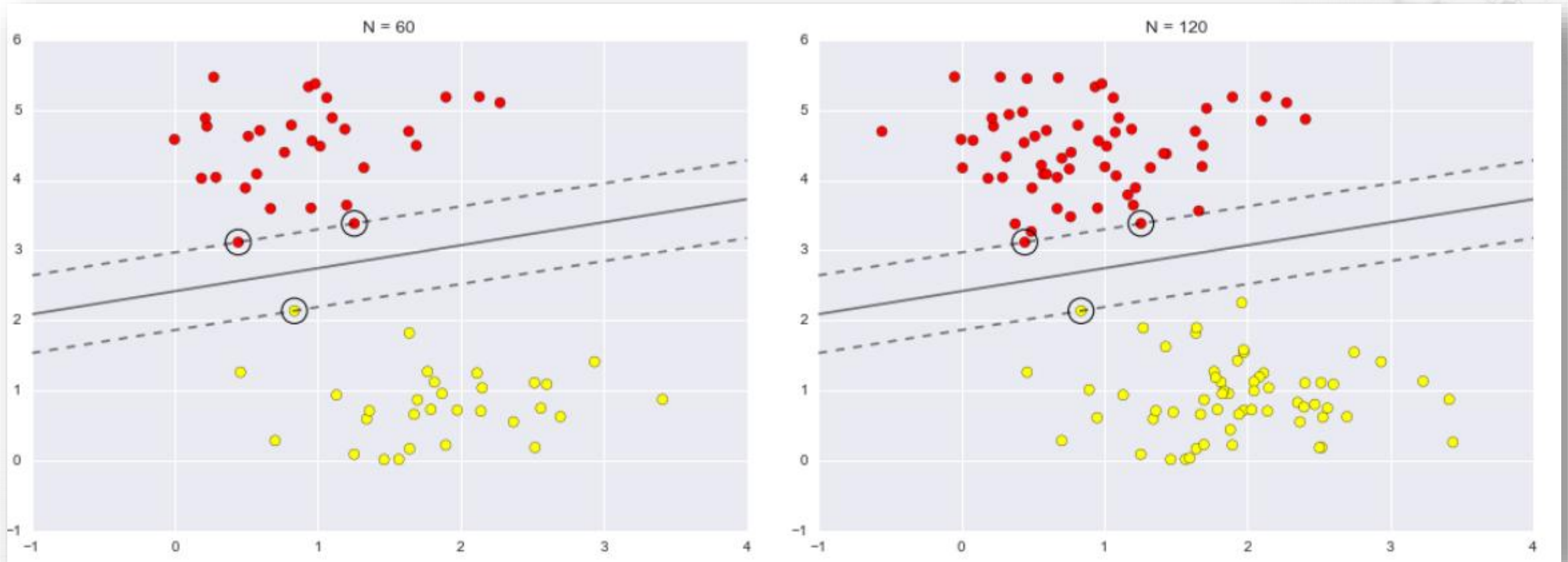
So what?

In höheren Dimensionen spricht man nicht mehr von Geraden, sondern Hyperebenen, die unsere Klassen trennen



Was denken Sie?
Wie sieht unser Modell in
höheren Dimensionen aus?

SVM: Support Vectors



0

So what?

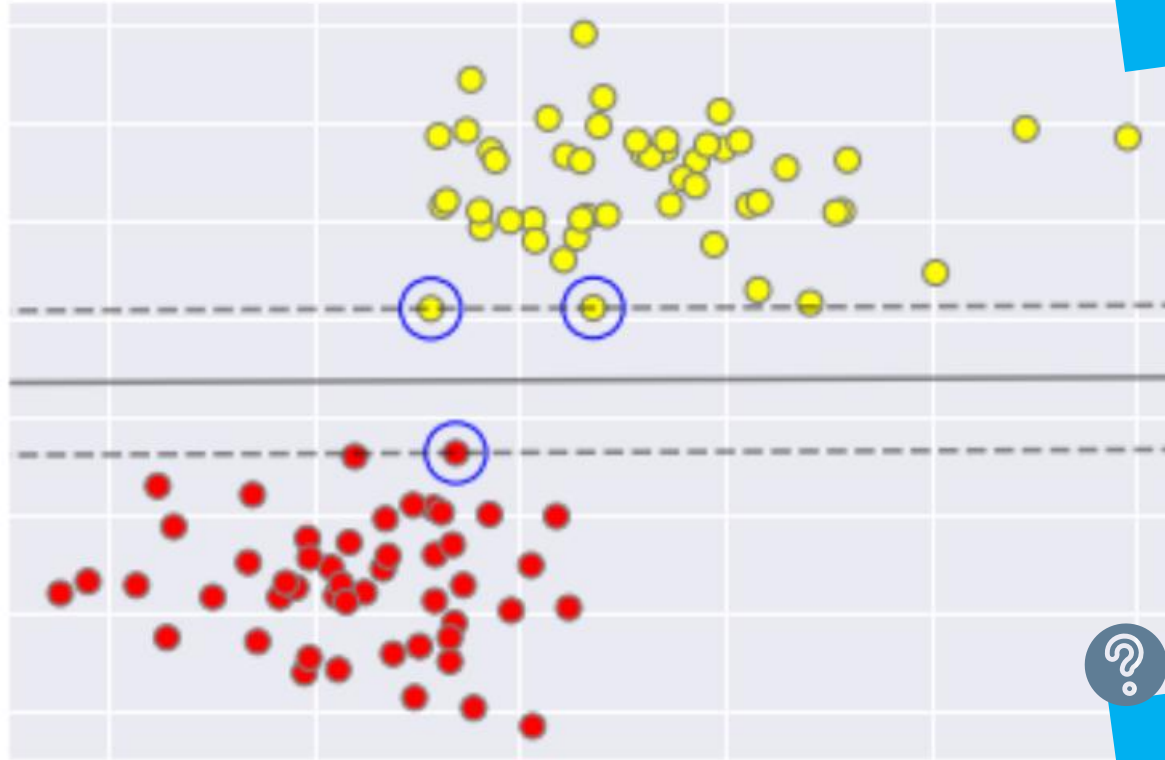
Diese Unempfindlichkeit des Modells auf entfernt liegende Punkt ist eine der Stärken dieser Art ML-Modelle

Wir als Klassifikatoren 2.0



Was denken Sie?

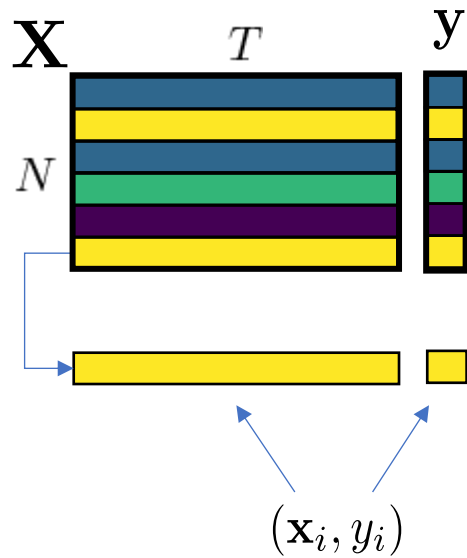
Wie würden Sie nun – als Klassifikator – eine optimale Trennlinie einzeichnen? (Zoom: zeichnen)



Was denken Sie?

Was sind die Support Vektoren? (Zoom: zeichnen)

SVM: Mathematik



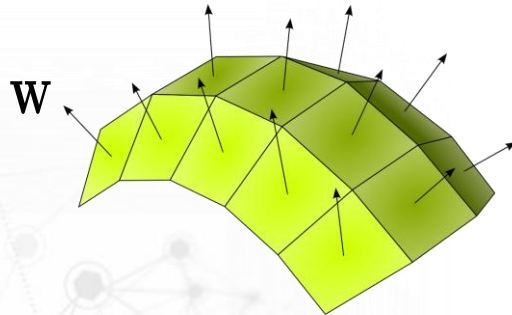
Uns liegt ein historischer Datensatz bzw. Trainingsdatensatz vor

$$\{(\mathbf{x}_i, y_i) | i = 1, \dots, N, y_i \in \{-1, 1\}\}$$

Anhand dessen wird unsere Hyperebene gelernt. Wir beschreiben unsere Hyperebene mittels des Normalenvektors \mathbf{W} durch den Ursprung

und des Abstandes vom Ursprung b , den man *Bias* nennt

Mit diesen beiden Informationen haben wir eindeutig eine Hyperebene im Raum definiert



Viele Normalenvektoren

SVM: Mathematik



Fenster Time!

Für die Datenpunkte auf der Hyperebene gilt

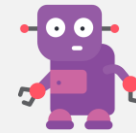
$$\langle \mathbf{w}, \mathbf{x}_i \rangle + b = 0$$

Für Punkte, die nicht auf der Hyperebene liegen ist diese Gleichung nicht 0, sondern positiv (Richtung \mathbf{w}) bzw. negativ (entgegengesetzt zu \mathbf{w})

Wenn also eine Hyperebene **alle Punkte** unserer Klassen voneinander **trennt**, dann weisen die Punkte einer Klasse alle entweder **positive** oder **negative** Vorzeichen auf

Für unsere Klassen gilt dann also

$$y_i = \text{sgn}(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)$$



Unser Modell!

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 0$$

Aus den unendlich vielen Hyperebenen wählt man nun die, die den größten *Margin* zwischen den beiden Klassen aufweist

Was denken Sie?
Wie drückt man das in einer Formel aus?



Was denken Sie?
Was gilt für Punkte, die nicht auf der Hyperebene liegen?

Was denken Sie?
Wie könnten wir das mit einer Ungleichung ausdrücken?

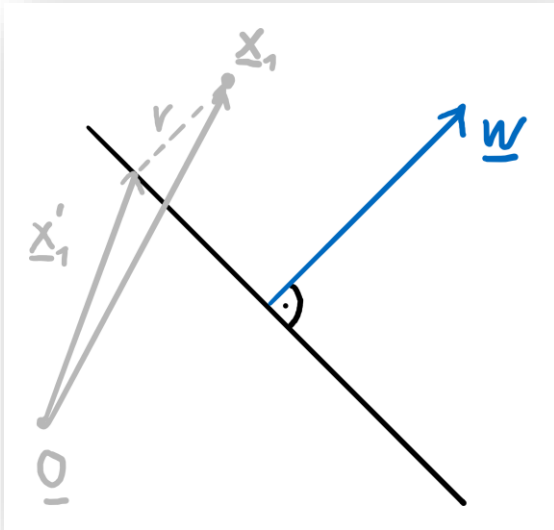
Was denken Sie?
Warum soll der Margin maximiert werden?

Was denken Sie?
Wie berechnet man also die Klassen/Labels neuer Daten?

Was denken Sie?
Was sucht also der Trainingsprozess?

SVM: Mathematik

$$\mathbf{x}_1 = \mathbf{x}'_1 + r \frac{\mathbf{w}}{\|\mathbf{w}\|}$$



Was bedeutet ein **maximaler Margin** mathematisch für uns?

Der Margin wird durch den **Abstand** zu den **nächsten** Punkten der Klassen zur Hyperebene definiert – wir brauchen also einen Abstand

Mit diesem Abstand r können wir nun unsere **Bedingung**, auf den jeweiligen Seiten der Hyperebene **nur Datenpunkte je einer Klasse** vorzufinden formulieren

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq r$$

Wir können nun ein **Optimierungsproblem** formulieren

$$\max_{\mathbf{w}, b} r$$

unter den **Bedingungen**

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq r \quad \|\mathbf{w}\| = 1 \quad r > 0$$

0

So what?

Wir versuchen nun sukzessive dieses Optimierungsproblem zu vereinfachen. Was heißt hier vereinfachen?

SVM: Mathematik



Was denken Sie?

Wieso heißen
diese Bedingungen
Hard Margin?

Der Normalenvektor sei nun **nicht mehr normalisiert**. Um das Optimierungsproblem *berechenbar* zu machen, müssen wir eine kleine Umformung machen: durch eine **Skalierung** der Daten auf die Bedingung

$$\langle \mathbf{w}, \mathbf{x}_a \rangle + b = 1$$

mit \mathbf{x}_a als der nächste Punkt zur Hyperebene (Support Vector) und den vorher definierten Abstand kann gezeigt werden, dass

$$r = \frac{1}{\|\mathbf{w}\|}$$

und für unsere Bedingung gilt dann

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$$

Damit reduziert sich unser Problem auf

$$\max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|} \quad \text{bzw.} \quad \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2$$

unter der Bedingung

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$$

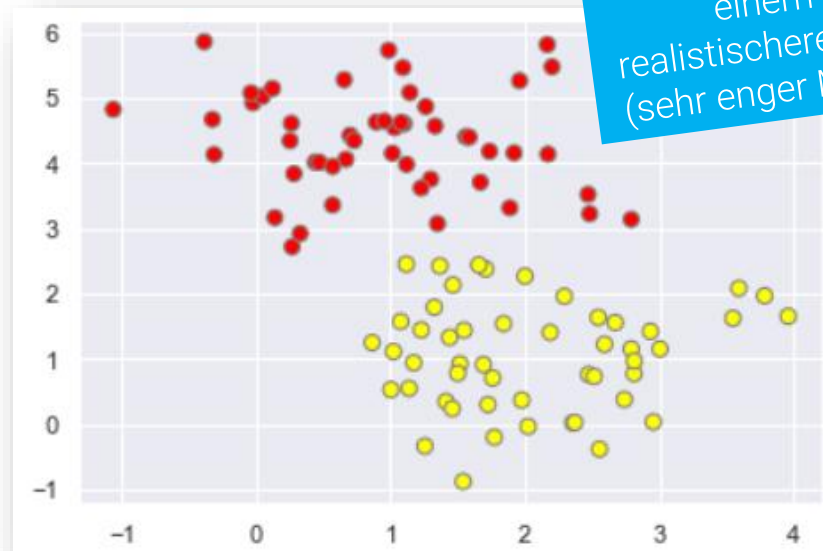


SVM: Hard- und Soft-Margin



Was denken Sie?

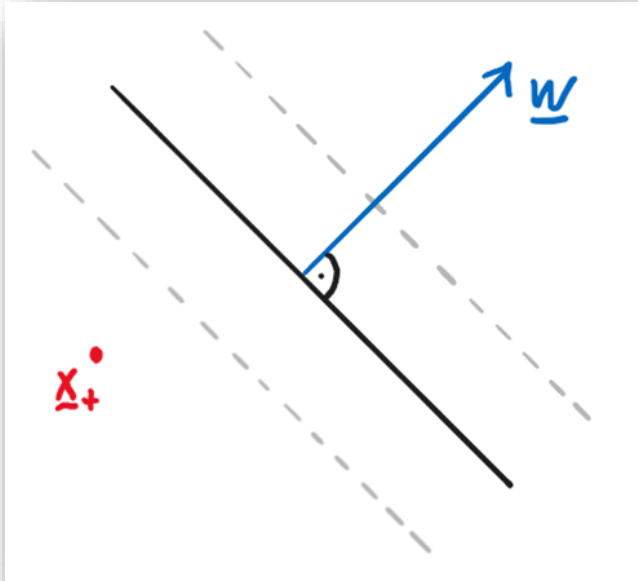
Was tun Sie in so einem – realistischeren – Fall (sehr enger Margin)?



```
1 X, y = make_blobs(n_samples=100, centers=2,  
2                   random_state=0, cluster_std=0.8)  
3 plt.scatter(X[:, 0], X[:, 1], c=y, s=50, cmap='autumn', edgecolors='gray');
```

- Bei SVMs hat man die Möglichkeit die Margins entweder starr bzw. **hard** oder weich bzw. **soft** zu definieren
- In anderen Worten: lässt man zu, dass Datenpunkte **innerhalb** des Margins auftauchen dürfen – Leitfrage: **wie viele** bzw. zu welchem Ausmaß?
- Hierzu gibt es einen Hyperparameter, der getuned werden muss – in `sklearn` ist es das Argument `c`
- **Große** Werte entsprechen einem **Hard**-, **kleine** Werte einem **Soft-Margin**
- Der **optimale** Wert von `c` hängt also vom Datensatz ab und muss via z.B. Cross-Validation bestimmt werden
→ Zuerst schauen wir, wo in der Mathematik dieses `C` zu finden ist

SVM: Soft-Margin: Mathematik



Was denken Sie?

Was ändern wir an unseren Formulierungen? Wie quantifizieren wir Verletzungen des Margins?

Wir führen eine sog. *Slack Variable* ein, die **bestraft** wenn ein Datenpunkt innerhalb des Margins bzw. auf der falschen Seite liegt – dies ermöglicht uns im Umkehrschluss: **Verletzungen des Margins werden zugelassen!**

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i$$

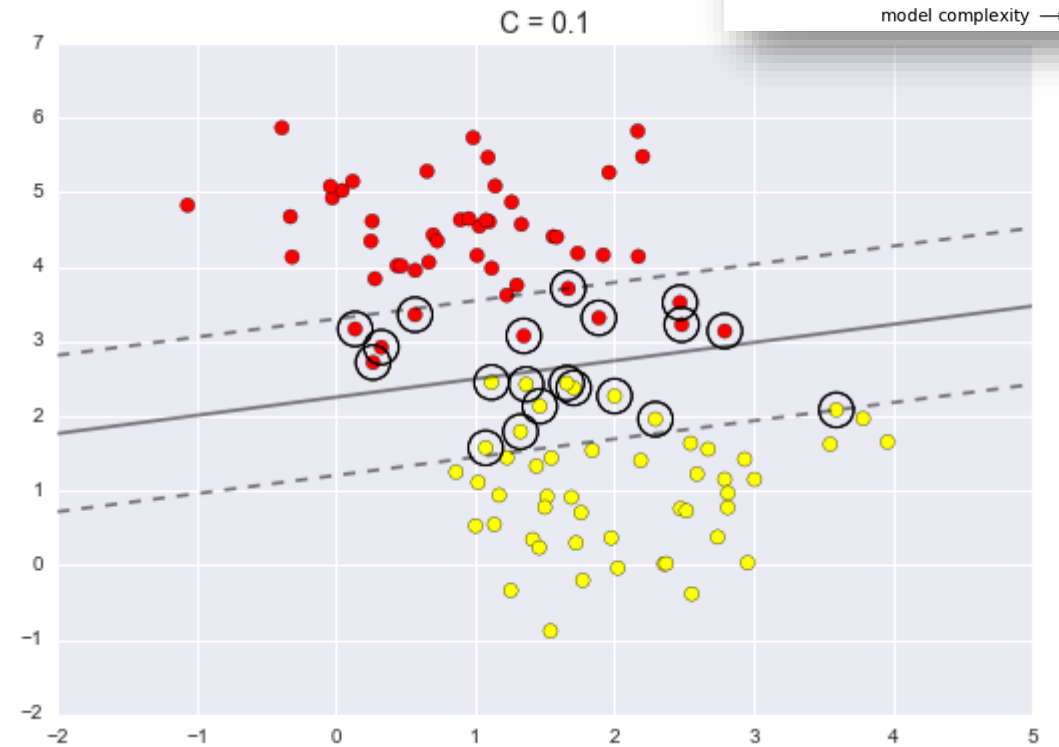
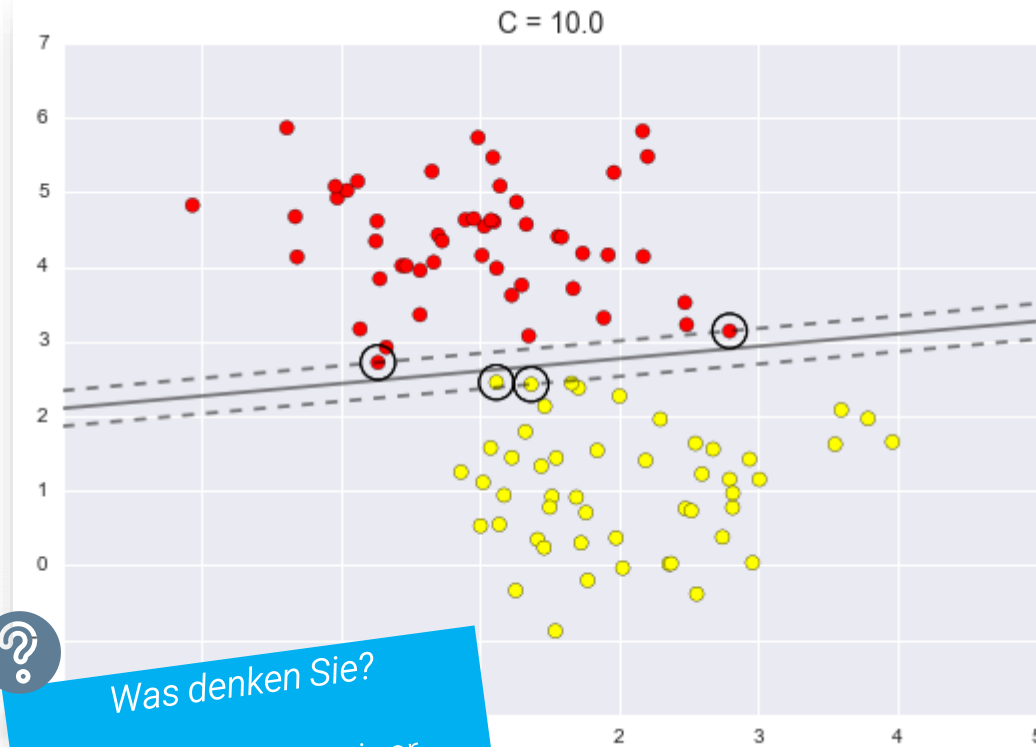
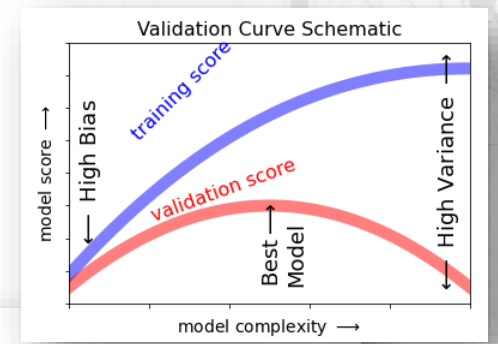
mit den Bedingungen

$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 - \xi_i \quad \text{und} \quad \xi_i \geq 0$$

Der Parameter C versucht also eine **Balance** bzw. einen **Trade-Off** zwischen „Größe des Margins“ und „Verletzung dessen“ herzustellen

Zu diesem Parameter sagt man **Regularisierungsparameter**

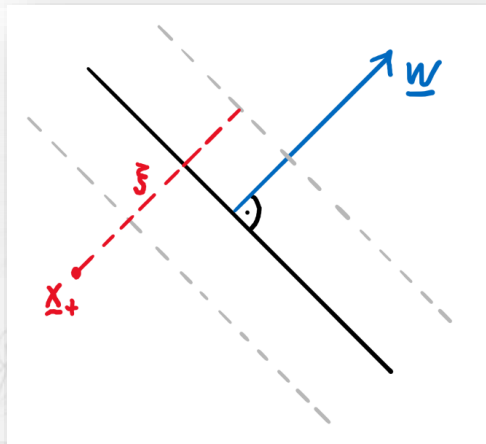
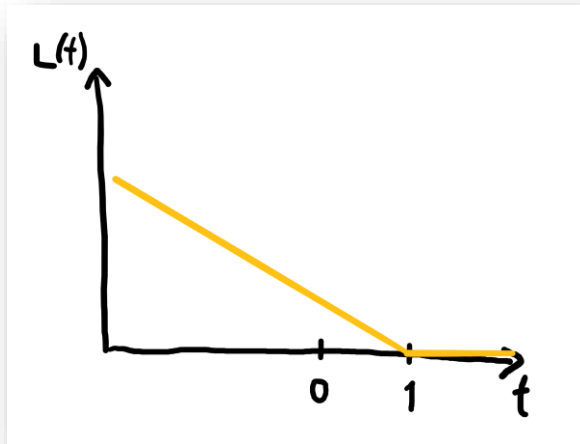
SVM: Hard- und Soft-Margin



Was denken Sie?

Was entspricht einer hohen, was einer niedrigen Modellkomplexität?

SVM: Optimierung mittels Loss Function



Die naheliegendste Formulierung einer Loss Function wäre die Accuracy – also: wie gut decken sich Vorhersage und Labels?

Diese ist jedoch ungeeignet, da es sich hierbei um ein sog. *kombinatorisches Optimierungsproblem* handelt

Eine *kontinuierliche Loss Function* ist also anzustreben – mögliche Formulierung: **Hinge Loss**

$$L(t) = \max\{0, 1 - t\} \quad \text{mit} \quad t = yf(\mathbf{x}) = y(\langle \mathbf{w}, \mathbf{x} \rangle + b)$$

Damit ergibt sich unser letztendlich **optimierbares** Problem zu

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \max\{0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\}$$

Hierzu sagt man auch **Regularisierung** – **Margin Maximierung** ist also Regularisierung

So what?

Durch diesen Hinge-Loss konnten wir die vorher eingeführte **Slack-Variable** wieder ersetzen und dadurch unser Optimierungsproblem vereinfachen.



Was denken Sie?

Wie würden Sie die Kostenfunktion aufstellen?



Was denken Sie?

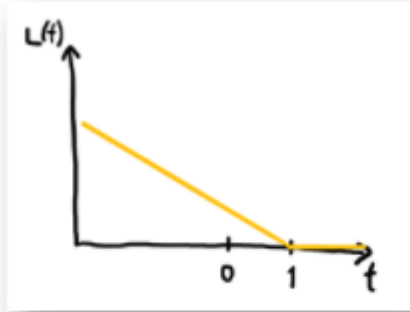
Wie schaut die math. Formulierung hierzu aus?



Was denken Sie?

Wie leitet man die max-Funktion ab?

Beispiel: SVM from Scratch: Loss Function



Wir haben gesehen, dass man durch ein paar mathematische Umformungen zu einem übersichtlichen Optimierungsproblem bzw. einer relativ überschaubaren Kostenfunktion gelangt. Wir wollen uns nun in diesem Beispiel damit beschäftigen die Loss Function

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \max\{0, 1 - y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)\}$$

als Funktion in Python abzubilden.



Was denken Sie?

Ohne es hier schon behandelt zu haben: haben Sie eine Vermutung wie man damit nun die optimalen Modellparameter findet?



So what?

Nun wissen wir wie die SVM als mathematisches Modell aussieht und welche Kostenfunktion minimiert werden soll. Das Werkzeug für die Minimierung – z.B. den sog. *Gradientenabstieg* – sehen wir uns in einer späteren Veranstaltung an.