CS5489 - Machine Learning

Lecture 3b - Support Vector Machines

Prof. Antoni B. Chan

Dept. of Computer Science, City University of Hong Kong

Outline

1. Discriminative classifiers
2. Logistic regression
3. **Support vector machines**

Support vector machines

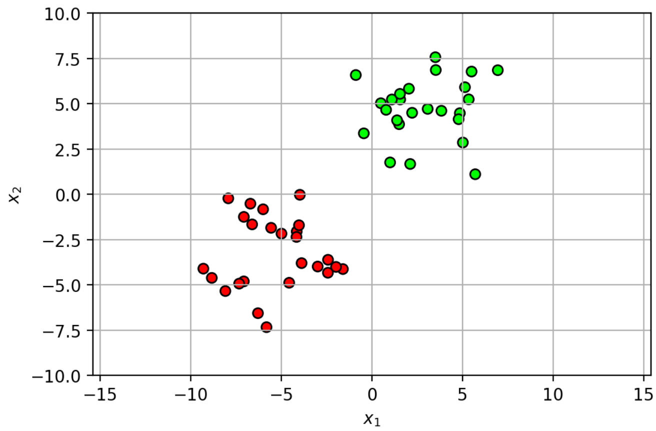
- With logistic regression we used a maximum-likelihood framework to learn the separating hyperplane.
- Let's consider a purely geometric approach...

Linearly-Separable Data

- For now, assume the training data is *linearly separable*
  - the two classes in the training data can be separated by a line (hyperplane)

In [3]:
```
lsdatafig
```

Out[3]:

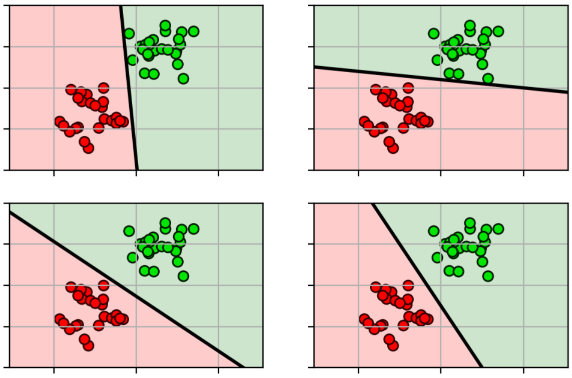


Which is the best separating line?

- there are many possible solutions...

In [6]:
```
seplinefig
```

Out[6]:

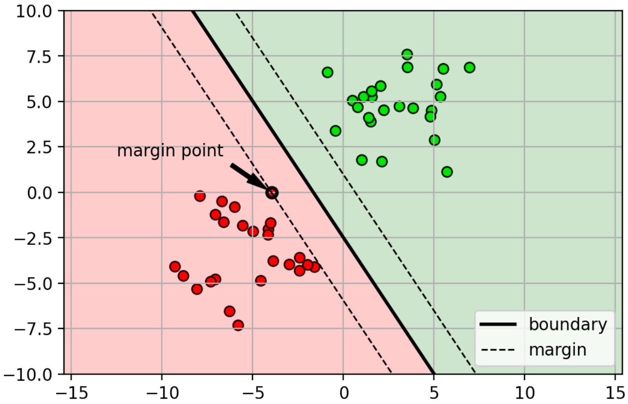


Maximum margin

- Define the space between the separating line and the closest point as the *margin*.
  - think of this space as the "amount of wiggle room" for accommodating errors in estimating $\mathbf{w}$.

In [9]:
```
margfig
```

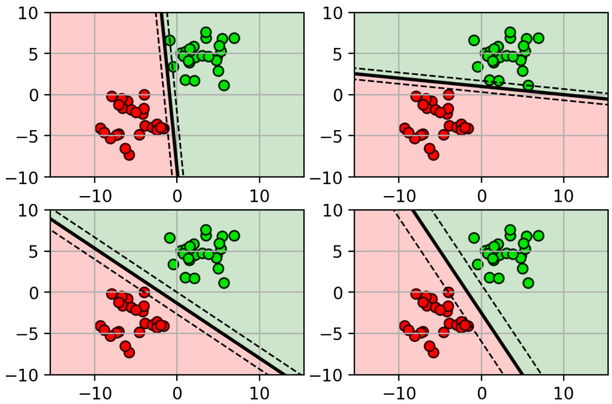
Out[9]:



- **Idea:** the best separating line is the one that *maximizes the margin*.
  - i.e., puts the most distance between the closest points and the decision boundary.

In [11]:
```
margfigs
```

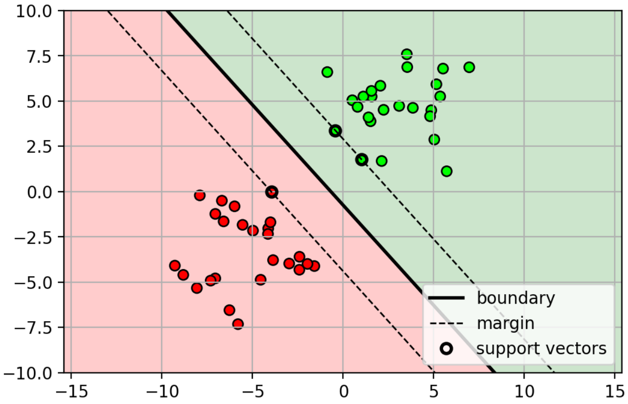
Out[11]:



- *the solution...*
  - by symmetry, there should be at least one margin point on each side of the boundary
    - if there is only one margin point, we could increase the margin distance by moving the boundary away from the margin point.
  - the points on the margins are called the **support vectors**
    - the points support (define) the margin

In [13]:
```
maxmfig
```

Out[13]:

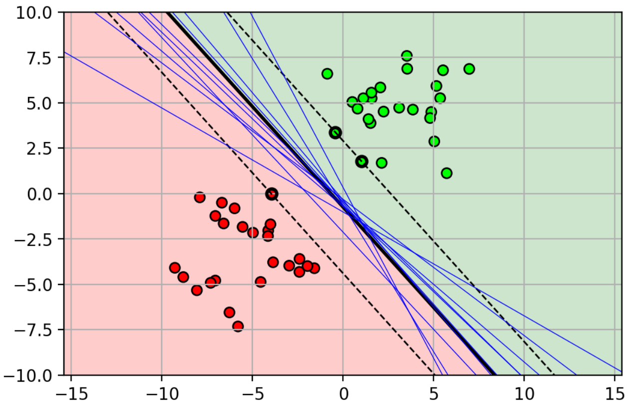


Why is maximizing the margin good?

- the true $\mathbf{w}$ is uncertain
  - maximizing the margin allows the most uncertainty (wiggle room) for $\mathbf{w}$, while keeping all the points correctly classified.

In [15]:
```
maxmfigw
```

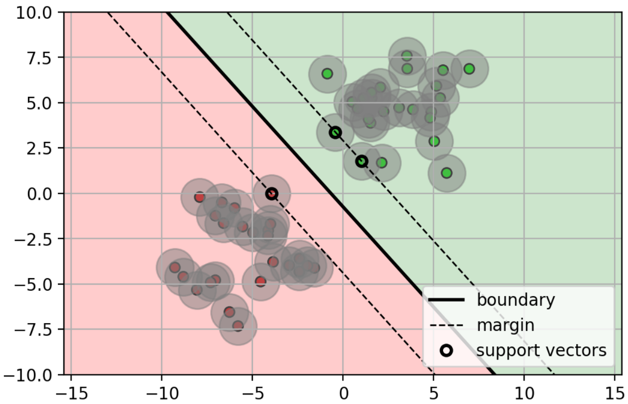
Out[15]:



- the data points are uncertain
  - maximizing the margin allows the most wiggle of the points, while keeping all the points correctly classified.

In [17]:
```
maxmfigp
```

Out[17]:



SVM Training

- Given a training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$.
- First define the margin distance:
  - Distance from a point $\mathbf{x}_i$ to hyperplane $\mathbf{w}$:

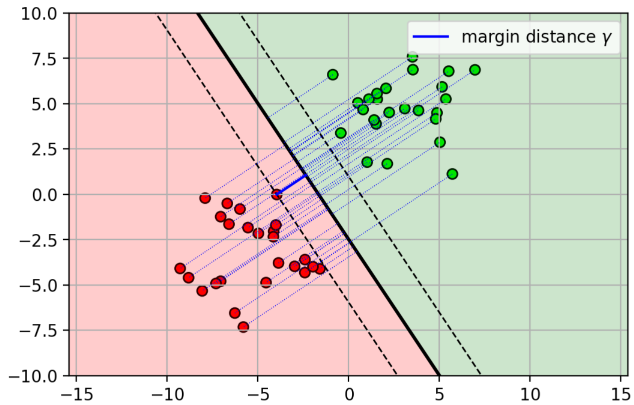
$$d_i = \frac{|f(\mathbf{x}_i)|}{||\mathbf{w}||}$$

- Margin distance is the minimum distance among all points:

$$\gamma = \min_i \frac{|f(\mathbf{x}_i)|}{||\mathbf{w}||}$$

In [20]:
```
dfig
```

Out[20]:



- The hyperplane $\mathbf{w}$ appears in both numerator and denominator!
  - $\gamma = \min_i \frac{|f(\mathbf{x}_i)|}{||\mathbf{w}||} = \min_i \frac{|\mathbf{w}^T\mathbf{x}_i + b|}{||\mathbf{w}||}$
- Changing the length of $\mathbf{w}$ won't affect the margin distance.
  - $\hat{\mathbf{w}} = a\mathbf{w}$, $\hat{b} = ab$
  - $\frac{|\hat{\mathbf{w}}^T\mathbf{x}_i + \hat{b}|}{||\hat{\mathbf{w}}||} = \frac{|a\mathbf{w}^T\mathbf{x}_i + ab|}{||a\mathbf{w}||} = \frac{|\mathbf{w}^T\mathbf{x}_i + b|}{||\mathbf{w}||}$
- Margin distance is determined by the direction of $\mathbf{w}$, not the length.

Normalization

- Since the length of $\mathbf{w}$ doesn't matter, we can assume a normalization for $\mathbf{w}$.
- Two possibilities:
  1. set denominator to 1: $||\mathbf{w}|| = 1$
     - $||\mathbf{w}||$ is a unit-norm vector
  2. set numerator to 1: $\min_i |f(\mathbf{x}_i)| = 1$
     - the point $\mathbf{x}_i$ on the margin has $f(\mathbf{x}_i) = 1$.
- Which is better?

SVM Optimization Problem

- Choose the 2nd option.
  - constraint: $\min_i |f(\mathbf{x}_i)| = 1$
  - margin: $\gamma = \frac{1}{||\mathbf{w}||}$
- Maximize the margin:

$$(\hat{\mathbf{w}}, b) = \operatorname*{argmax}_{\mathbf{w},b} \frac{1}{||\mathbf{w}||} \quad \text{s.t.} \ \min_i |f(\mathbf{x}_i)| = 1$$

- Invert the objective to turn into a minimization problem

$$(\hat{\mathbf{w}}, b) = \operatorname*{argmin}_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2 \quad \text{s.t.} \ \min_i |f(\mathbf{x}_i)| = 1$$

- Note: if the points are correctly classified...
  - for points in class $y_i = 1$, then $f(\mathbf{x}_i) > 0$.
  - for points in class $y_i = -1$, then $f(\mathbf{x}_i) < 0$.
- Thus, the constraint: $\min_i |f(\mathbf{x}_i)| = 1,$
  - can be rewritten: $\min_i y_i f(\mathbf{x}_i) = 1$

- Closer look:
  - original constraint: $\min_i y_i f(\mathbf{x}_i) = 1$
    - the minimum over all i is 1.
  - equivalent constraint: $y_i f(\mathbf{x}_i) \geq 1, \forall i$
    - suppose $y_i f(\mathbf{x}_i) = y_i(\mathbf{w}^T \mathbf{x}_i + b) > 1, \forall i$.
    - since we are minimizing $||\mathbf{w}||$, $\mathbf{w}$ will shrink so that at least one $\mathbf{x}_i$ will have $y_i f(\mathbf{x}_i) = 1$.
      - $(\mathbf{w}, b) \Rightarrow (\delta \mathbf{w}, \delta b)$, where $y_i(\delta \mathbf{w}^T \mathbf{x}_i + \delta b) \geq 1, \forall i$

SVM Training Objective

- given a training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$, optimize:

$$\begin{aligned} &\operatorname*{argmin}_{\mathbf{w},b} \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ &\text{s.t. } y_i f(\mathbf{x}_i) \geq 1, \quad \forall i \end{aligned}$$

  - the objective minimizes the inverse of the margin distance, i.e., maximizes the margin.
  - the inequality constraints ensure that all points are either on or outside of the margin.
    - a point on the margin has $y_i f(\mathbf{x}_i) = 1$.
    - also written as $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$.

SVM Prediction

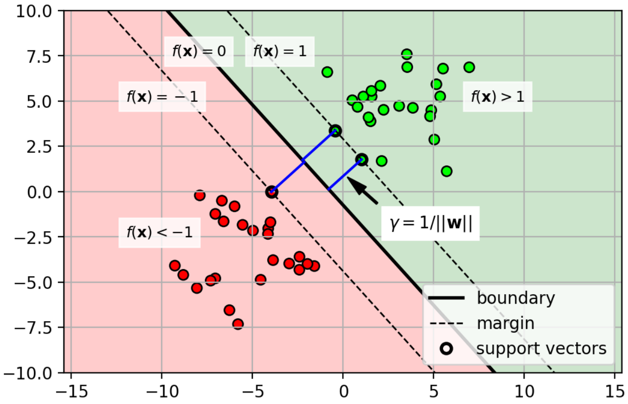
- given a new data point $\mathbf{x}_*$, use sign of linear function to predict class
  - $y_* = \operatorname{sign} f(\mathbf{x}_*) = \operatorname{sign}(\mathbf{w}^T \mathbf{x}_* + b)$

Example

- Regions defined by $f(\mathbf{x})$:
  - $f(\mathbf{x}) = 0$ -- decision boundary
  - $f(\mathbf{x}) = \pm 1$ -- positive and negative margins
  - $f(\mathbf{x}) > 1$ -- points correctly classified as class 1
  - $f(\mathbf{x}) < -1$ -- points correctly classified as class -1

In [22]:
```
egsvmfig
```

Out[22]:



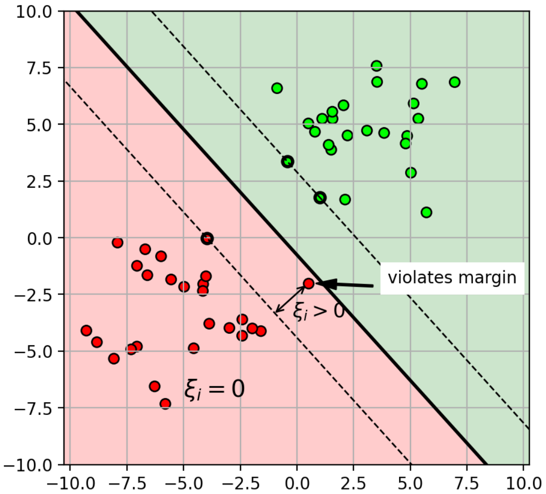
What about non-separable data?

- use the same linear classifier
  - allow some training samples to **violate** the margin
    - i.e., are inside the margin (or even mis-classified)

- Define "slack" variable $\xi_i \geq 0$
  - $\xi_i = 0$ means sample is outside of margin area (no slack)
  - $\xi_i > 0$ means sample is inside of margin area (slack)

In [24]:
```
slackfig
```

Out[24]:



- constraint now includes slack variable
  - $y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \forall i$
- Two possibilities:
  - $\xi_i = 0$, then sample $\mathbf{x}_i$ has normal margin constraint: $y_i f(\mathbf{x}_i) \geq 1$
  - $\xi_i > 0$, then sample $\mathbf{x}_i$ is inside margin: $y_i f(\mathbf{x}_i) = 1 - \xi_i$
    - Note that: $y_i f(\mathbf{x}_i) < 1$, inside margin.

Soft-SVM optimization problem

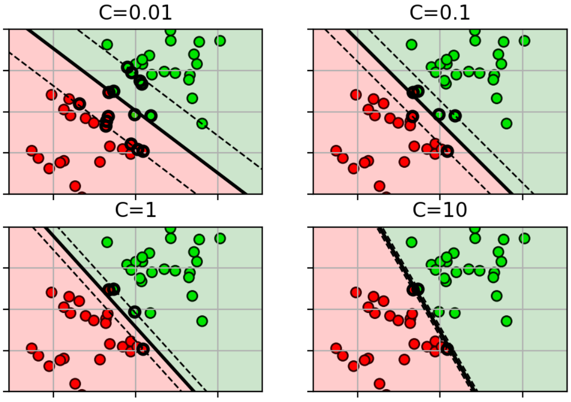
- Penalize each training sample that violates the margin by summing over ξ_i.
  - penalty controlled by hyperparameter C.
    - smaller value means allow more violations (less penalty)
    - larger value means don't allow violations (more penalty)

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^N \xi_i$$
$$\text{s.t. } y_i f(\mathbf{x}_i) \geq 1 - \xi_i, \quad \forall i$$

- Example with different C.

In [26]:
```
Cmargfigs
```

Out[26]:



Loss function

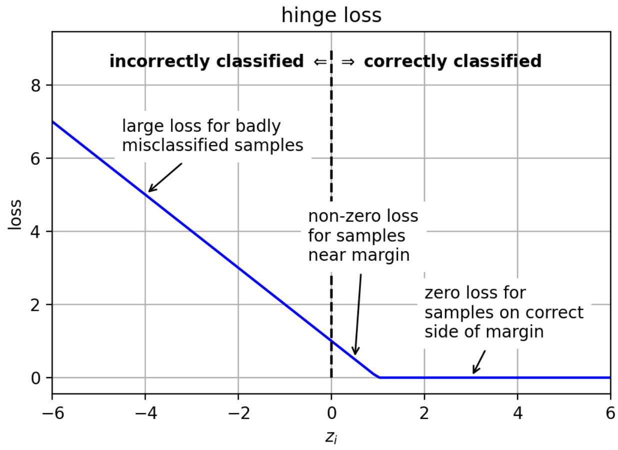
- After some massaging, the objective function is:

$$\underset{\mathbf{w},b}{\operatorname{argmin}} \frac{1}{C}\mathbf{w}^T\mathbf{w} + \sum_{i=1}^N \max(0, 1 - y_i f(\mathbf{x}_i))$$

- hinge loss function: $L(z_i) = \max(0, 1 - z_i)$
  - Note: $\max(a, b)$ returns whichever value (a or b) is largest.

In [28]:
```
lossfig
```

Out[28]:



Example: Iris Data

In [29]:
```
# load iris data each row is (petal length, sepal width, class)
irisdata = loadtxt('iris2.csv', delimiter=',', skiprows=1)

X = irisdata[:,0:2]  # the first two columns are features (petal length, sepal width)
Y = irisdata[:,2]    # the third column is the class label (versicolor=1, virginica=2)

print(X.shape)
```

```
(100, 2)
```

In [30]:
```python
# randomly split data into 50% train and 50% test set
trainX, testX, trainY, testY = \
  model_selection.train_test_split(X, Y,
  train_size=0.5, test_size=0.5, random_state=4487)

print(trainX.shape)
print(testX.shape)
```

```
(50, 2)
(50, 2)
```

In [31]:
```python
# fit the SVM using all the data and the best C
clf = svm.SVC(kernel='linear', C=2)
clf.fit(trainX, trainY)

# get line parameters
w = clf.coef_[0]
b = clf.intercept_[0]
print(w)
print(b)
```

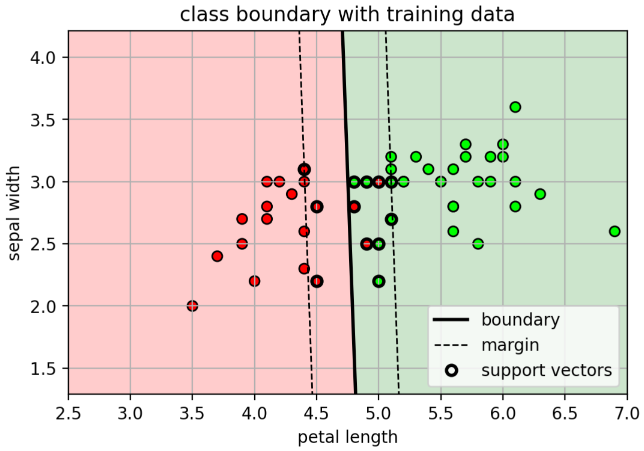
```
[2.87200943 0.10399865]
-13.95923965217775
```

In [32]:
```python
# indices of data points that are support vectors (on or inside the margin)
clf.support_
```

Out[32]:
```
array([ 0,  7, 12, 31, 36, 41, 13, 20, 22, 25, 33, 46], dtype=int32)
```

In [34]:
```python
svmfig
```

Out[34]:



- SVM doesn't have it's own dedicated cross-validation function
- Use the `GridSearchCV` to run cross-validation for a list of parameters
  - calculate average accuracy for each parameter
  - select parameter with average highest accuracy, retrain model with all data
  - Speed up: each parameter can be trained/tested separately, specify number of parallel jobs using `n_jobs`

In [35]:
```python
# setup the list of parameters to try
paramgrid = {'C': logspace(-3,3,13)}
print(paramgrid)

# setup the cross-validation object
# pass the SVM object, parameter grid, and number of CV folds
# set number of parallel jobs to -1 (use all cores)
svmcv = model_selection.GridSearchCV(svm.SVC(kernel='linear'), paramgrid, cv=5,
                                     n_jobs=-1, verbose=True)

# run cross-validation (train for each split)
svmcv.fit(trainX, trainY);
```

```
{'C': array([1.00000000e-03, 3.16227766e-03, 1.00000000e-02, 3.16227766e-02,
       1.00000000e-01, 3.16227766e-01, 1.00000000e+00, 3.16227766e+00,
```

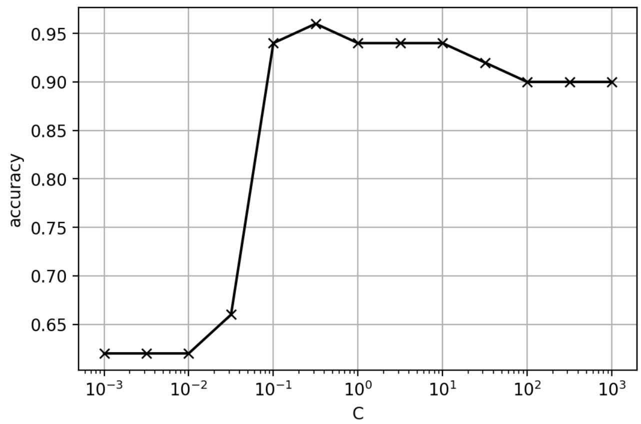
```
       1.00000000e+01, 3.16227766e+01, 1.00000000e+02, 3.16227766e+02,
       1.00000000e+03])}
Fitting 5 folds for each of 13 candidates, totalling 65 fits
```

In [36]:
```
# show the test error for each parameter set
for m,p in zip(svmcv.cv_results_['mean_test_score'], svmcv.cv_results_['params']):
    print("mean={:.4f} {}".format(m,p))
```

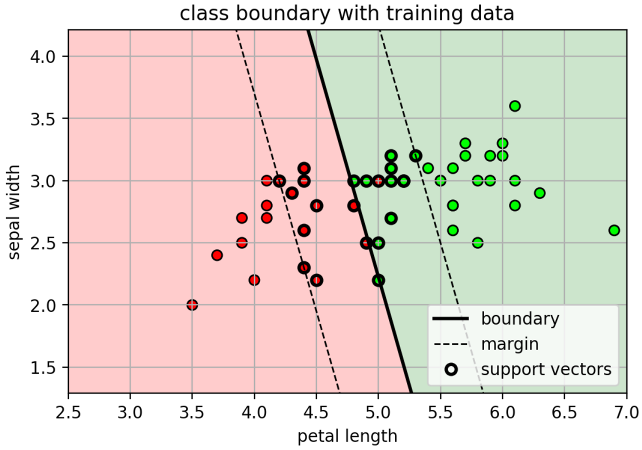
```
mean=0.6200 {'C': 0.001}
mean=0.6200 {'C': 0.0031622776601683794}
mean=0.6200 {'C': 0.01}
mean=0.6600 {'C': 0.03162277660168379}
mean=0.9400 {'C': 0.1}
mean=0.9600 {'C': 0.31622776601683794}
mean=0.9400 {'C': 1.0}
mean=0.9400 {'C': 3.1622776601683795}
mean=0.9400 {'C': 10.0}
mean=0.9200 {'C': 31.622776601683793}
mean=0.9000 {'C': 100.0}
mean=0.9000 {'C': 316.22776601683796}
mean=0.9000 {'C': 1000.0}
```

In [37]:
```
# make a plot
allC      = [p['C'] for p in svmcv.cv_results_['params']]
allscores = svmcv.cv_results_['mean_test_score']

plt.figure()
plt.semilogx(allC, allscores, 'kx-')
plt.xlabel('C'); plt.ylabel('accuracy')
plt.grid(True)

# view best results and best retrained estimator
print(svmcv.best_params_)
print(svmcv.best_score_)
print(svmcv.best_estimator_)
```

```
{'C': 0.31622776601683794}
0.96
SVC(C=0.31622776601683794, kernel='linear')
```



In [38]:
```
plt.figure()
plot_svm(svmcv.best_estimator_, axbox, mycmap)
plt.scatter(trainX[:,0], trainX[:,1], c=trainY, cmap=mycmap, edgecolors='k')
plt.xlabel('petal length'); plt.ylabel('sepal width')
plt.title('class boundary with training data');
```

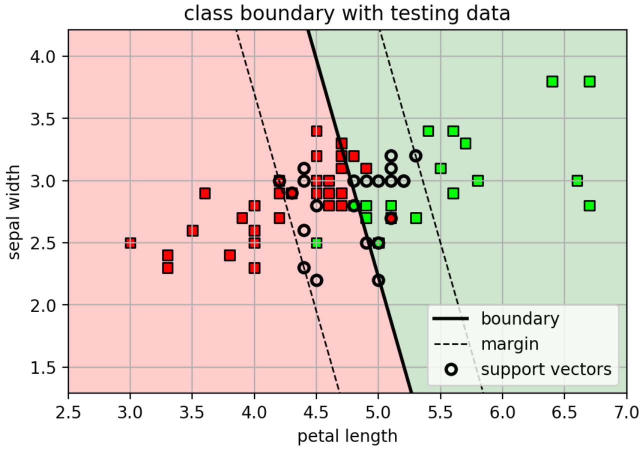


In [39]:
```
# Directly use svmcv to make predictions
predY = svmcv.predict(testX)

acc = metrics.accuracy_score(testY, predY)
print("test accuracy = " + str(acc))
```

```
test accuracy = 0.88
```

In [41]:
```
tsvmfig
```

Out[41]:

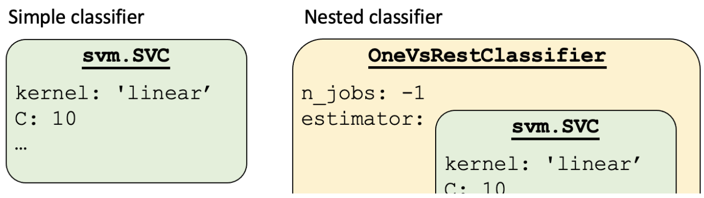


Multi-class SVM

- In sklearn, `svm.SVC` implements "1-vs-1" multi-class classification.
  - Train binary classifiers on all pairs of classes.
    - 3-class Example: 1 vs 2, 1 vs 3, 2 vs 3
  - To label a sample, pick the class with the most votes among the binary classifiers.
- Problem:
  - 1v1 classification is very slow when there are a large number of classes.
    - if there are C classes, need to train $C(C-1)/2$ binary classifiers!

1-vs-all SVM

- Use the `multiclass.OneVsRestClassifier` to build a 1-vs-all classifier from any binary classifier.
  - pass it the binary classifier as the base classifier.
- For `GridSearchCV`, the binary SVM is embedded inside the 1-vs-all wrapper class.
  - use `'estimator__C'` as the parameter label for `C` in the SVM.
  - notation `A__B` means the cross-validated parameter `B` is nested in parameter `A`.



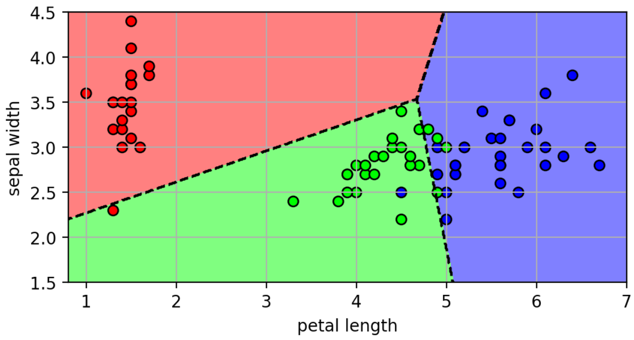
In [43]:
```
msvm = multiclass.OneVsRestClassifier(svm.SVC(kernel='linear'))

# setup the parameters and run CV
paramgrid = {'estimator__C': logspace(-3,3,13)}
msvmcv = model_selection.GridSearchCV(msvm, paramgrid, cv=5, n_jobs=-1, verbose=True)
msvmcv.fit(trainX, trainY)
print(msvmcv.best_params_)
```

```
Fitting 5 folds for each of 13 candidates, totalling 65 fits
{'estimator__C': 10.0}
```

3-class decision boundaries

In [45]:
```
svm3fig
```

Out[45]:

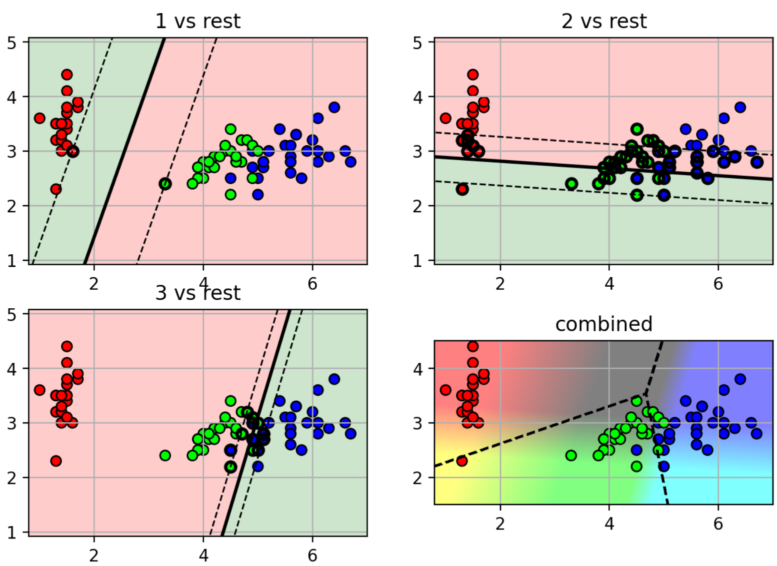


Decision boundaries for each binary classifier

In [47]:
```
for bclf in msvmcv.best_estimator_.estimators_:
    print(bclf.coef_)
bfig
```

```
[[-1.04615354  0.36923066]]
[[-0.14902716 -2.23820701]]
[[ 4.44425028 -1.33309667]]
```

Out[47]:



SVM Summary

- **Classifier:**
  - linear function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$
  - given new sample $\mathbf{x}_*$, predict $y_* = \text{sign}(f(\mathbf{x}_*))$.
- **Training:**
  - Maximize the margin of the training data.
    - i.e., maximize the separation between the points and the decision boundary.
  - Allow some training samples to violate the margin.
    - Use cross-validation to pick the hyperparameter C.

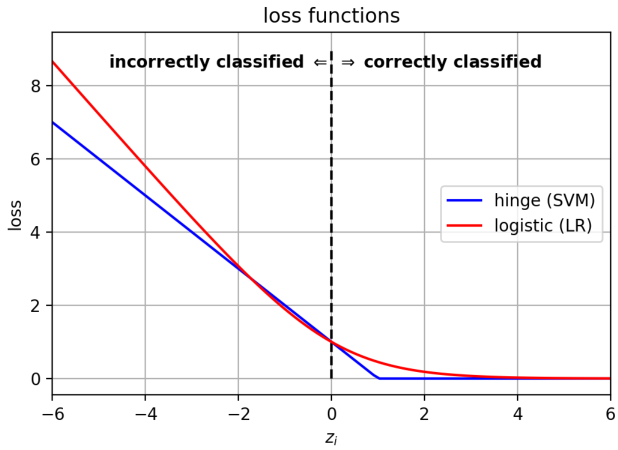
Summary

- **Linear classifiers:**
  - separate the data using a linear surface (hyperplane).
  - $y = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$
- **Two formulations:**
  - logistic regression - maximize the probability of the data
  - support vector machine - maximize the margin of the hyperplane

- **Loss functions**
  - SVM - ensure a margin of 1 between boundary and closest point
  - LR - push the classification boundary as far as possible from all points

In [49]:
```
lossfig
```

Out[49]:



- **Advantages:**
  - SVM works well on high-dimensional features (d large), and has low generalization error.
  - LR has output probabilities.
- **Disadvantages:**
  - decision surface can only be linear!
    - Next lecture we will see how to deal with non-linear decision surfaces.