Lecture 8
Support Vector Machines

EE-UY 4563/EL-GY 9123: INTRODUCTION TO MACHINE LEARNING

PROF. SUNDEEP RANGAN, WITH MODIFICATION BY YAO WANG

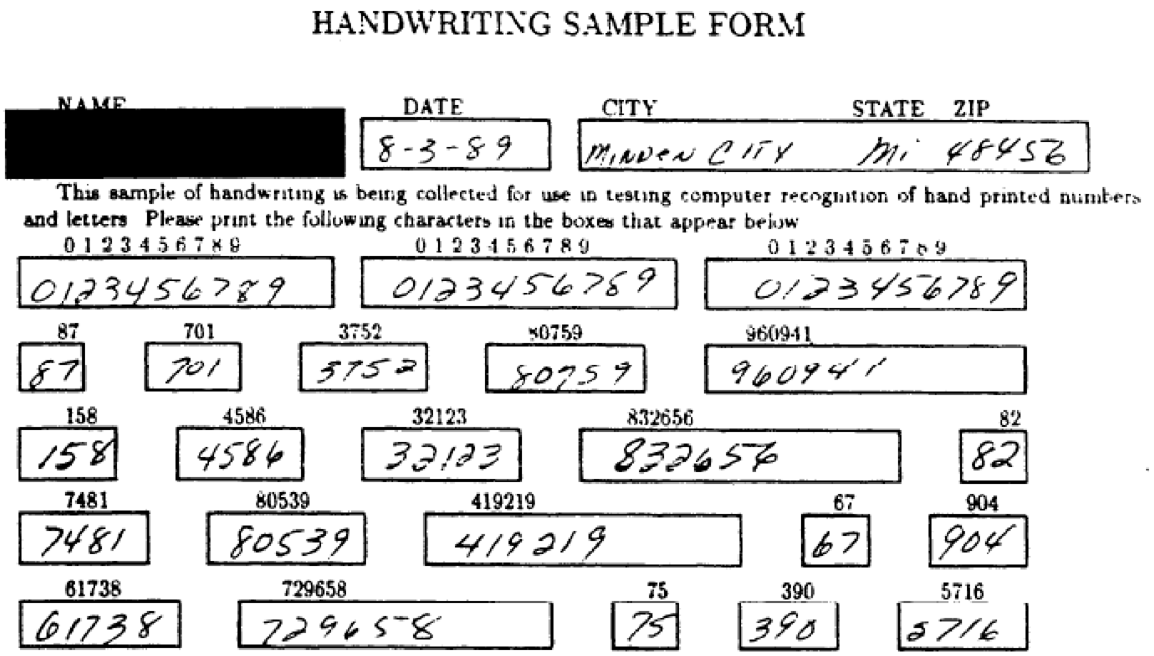
Learning Objectives

- ❑Interpret weights in linear classification of images
- ❑Define the margin in linear classification
- ❑Describe the SVM classification problem.
- ❑Write equations for solutions of constrained optimization using the Lagrangian.
- ❑Describe a kernel SVM problem
- ❑Select SVM parameters from cross-validation

Outline

- Motivating example: Recognizing handwritten digits
  - Why logistic regression doesn't work well.
- Maximum margin classifiers
- Support vector machines
- Constrained optimization
- Kernel trick

MNIST Digit Classification



HANDWRITING SAMPLE FORM

NAME [redacted] DATE 8-3-89 CITY MINDEN CITY STATE MI ZIP 48456

This sample of handwriting is being collected for use in testing computer recognition of hand printed numbers and letters. Please print the following characters in the boxes that appear below.

0 1 2 3 4 5 6 7 8 9 — 0123456789
0 1 2 3 4 5 6 7 8 9 — 0123456789
0 1 2 3 4 5 6 7 8 9 — 0123456789

87 — 87; 701 — 701; 3752 — 3752; 80759 — 80759; 960941 — 960941

158 — 158; 4586 — 4586; 32123 — 32123; 832656 — 832656; 82 — 82

7481 — 7481; 80539 — 80539; 419219 — 419219; 67 — 67; 904 — 904

61738 — 61738; 729658 — 729658; 75 — 75; 390 — 390; 5716 — 5716

From Patrick J. Grother, NIST Special Database, 1995

- Problem: Recognize hand-written digits
- Original problem:
  - Census forms
  - Automated processing
- Classic machine learning problem
- Benchmark

A Widely-Used Benchmark

Classifiers [edit]

This is a table of some of the machine learning methods used on the database and their error rates, by type of classifier:

Type	Classifier	Distortion	Preprocessing	Error rate (%)
Linear classifier	Pairwise linear classifier	None	Deskewing	7.6[9]
K-Nearest Neighbors	K-NN with non-linear deformation (P2DHMDM)	None	Shiftable edges	0.52[14]
Boosted Stumps	Product of stumps on Haar features	None	Haar features	0.87[15]
Non-Linear Classifier	40 PCA + quadratic classifier	None	None	3.3[9]
Support vector machine	Virtual SVM, deg-9 poly, 2-pixel jittered	None	Deskewing	0.56[16]
Neural network	2-layer 784-800-10	None	None	1.6[17]
Neural network	2-layer 784-800-10	elastic distortions	None	0.7[17]
Deep neural network	6-layer 784-2500-2000-1500-1000-500-10	elastic distortions	None	0.35[18]
Convolutional neural network	Committee of 35 conv. net, 1-20-P-40-P-150-10	elastic distortions	Width normalizations	0.23[8]

- We will look at SVM today
- Not the best algorithm
- But quite good
- ...and illustrates the main points

Downloading MNIST

```python
import tensorflow as tf

(Xtr,ytr),(Xts,yts) = tf.keras.datasets.mnist.load_data()

print('Xtr shape: %s' % str(Xtr.shape))
print('Xts shape: %s' % str(Xts.shape))

ntr = Xtr.shape[0]
nts = Xts.shape[0]
nrow = Xtr.shape[1]
ncol = Xtr.shape[2]
```

```
Xtr shape: (60000, 28, 28)
Xts shape: (10000, 28, 28)
```

- MNIST data is available in many sources
- Note: It has been removed from sklearn
- Tensorflow version:
  - 60000 training samples
  - 10000 test samples
- Each sample is a 28 x 28 images
- Grayscale: Pixel values $\in \{0,1,\ldots,255\}$
  - 0 = Black and
  - 255 = White

Matrix and Vector Representation

- For this demo, we reshape data from $N \times 28 \times 28$ to $N \times 784$
- But, you can easily go back and forth
- Also, scale the pixel values from -1 to 1

```
npix = nrow*ncol
Xtr = 2*(Xtr/255 - 0.5)
Xtr = Xtr.reshape((ntr,npix))

Xts = 2*(Xts/255 - 0.5)
Xts = Xts.reshape((nts,npix))
```

28

28

xsq

```
x = xsq.ravel()
```

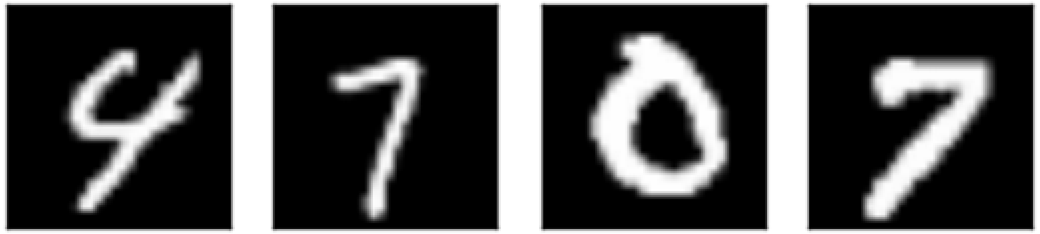
```
xsq = x.reshape((28,28))
```

$784 = 28^2$

$$S = \mathrm{Mat}(x) = \begin{bmatrix} s_{11} & \cdots & s_{1,28} \\ \vdots & \vdots & \vdots \\ s_{28,1} & \cdots & s_{28,28} \end{bmatrix}$$

$$x = \mathrm{vec}(S) = [x_1 \quad \cdots \quad x_{784}]$$

Displaying Images in Python



4 random images in the dataset

A human can classify these easily

```
def plt_digit(x):
    nrow = 28
    ncol = 28
    xsq = x.reshape((nrow,ncol))
    plt.imshow(xsq,  cmap='Greys_r')
    plt.xticks([])
    plt.yticks([])

# Convert data to a matrix
X = mnist.data
y = mnist.target

# Select random digits
nplt = 4
nsamp = X.shape[0]
Iperm = np.random.permutation(nsamp)

# Plot the images using the subplot command
for i in range(nplt):
    ind = Iperm[i]
    plt.subplot(1,nplt,i+1)
    plt_digit(X[ind,:])
```

Key command

Sample permutation is necessary for this dataset, as the original data is ordered by digits

Try a Logistic Classifier

```
ntr1 = 5000
Xtr1 = Xtr[Iperm[:ntr1],:]
ytr1 = ytr[Iperm[:ntr1]]
```

- Train on 5000 samples
  - To reduce training time.
  - In practice want to train with ~40k
- Select correct solver (lbfgs)
  - Others can be very slow. Even this will take minutes

```
from sklearn import linear_model
logreg = linear_model.LogisticRegression(verbose=10, solver='lbfgs',\
                                          multi_class='multinomial',max_iter=500)
logreg.fit(Xtr1,ytr1)
```

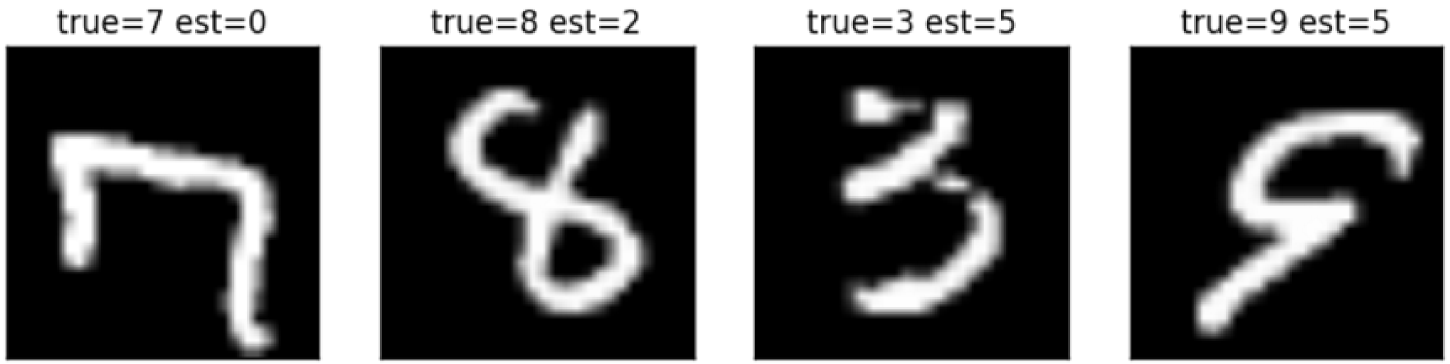
```
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:758:
e. Increase the number of iterations.
  "of iterations.", ConvergenceWarning)
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:  1.3min remaining:    0.0s
[Parallel(n_jobs=1)]: Done   1 out of   1 | elapsed:  1.3min finished
```

Performance

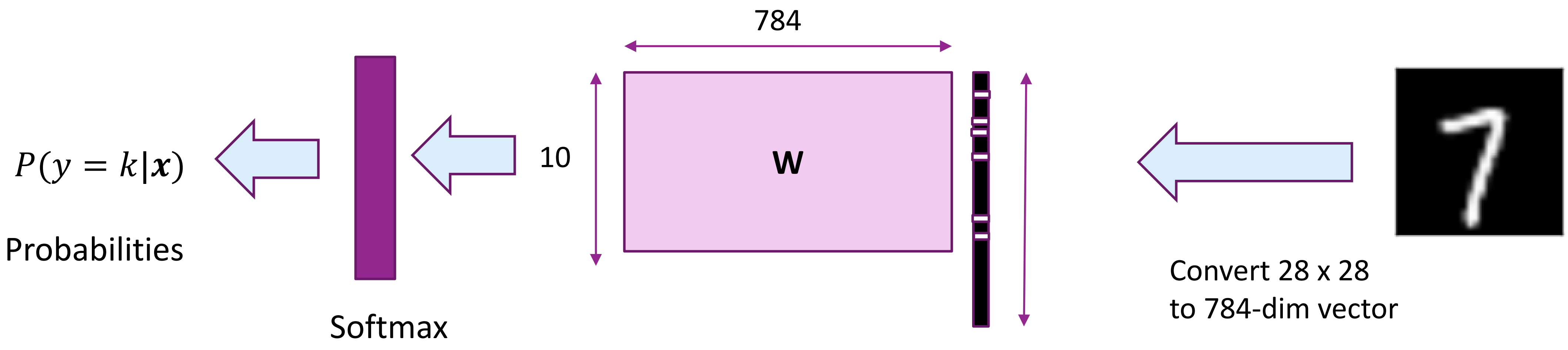
- Accuracy = 89%. Very bad
- Some of the errors seem like they should have been easy to spot
- What went wrong?

```
nts1 = 5000
Iperm_ts = np.random.permutation(nts)
Xts1 = Xts[Iperm_ts[:nts1],:]
yts1 = yts[Iperm_ts[:nts1]]
yhat = logreg.predict(Xts1)
acc = np.mean(yhat == yts1)
print('Accuaracy = {0:f}'.format(acc))
```

```
Accuaracy = 0.891000
```



Recap: Logistic Classifier



- Will select $\hat{y} = \arg\max_k P(y = k|x) = \arg\max_k z_k$
  - Output z_k which is largest
- When is z_k large?

Interpreting the Logistic Classifier Weights

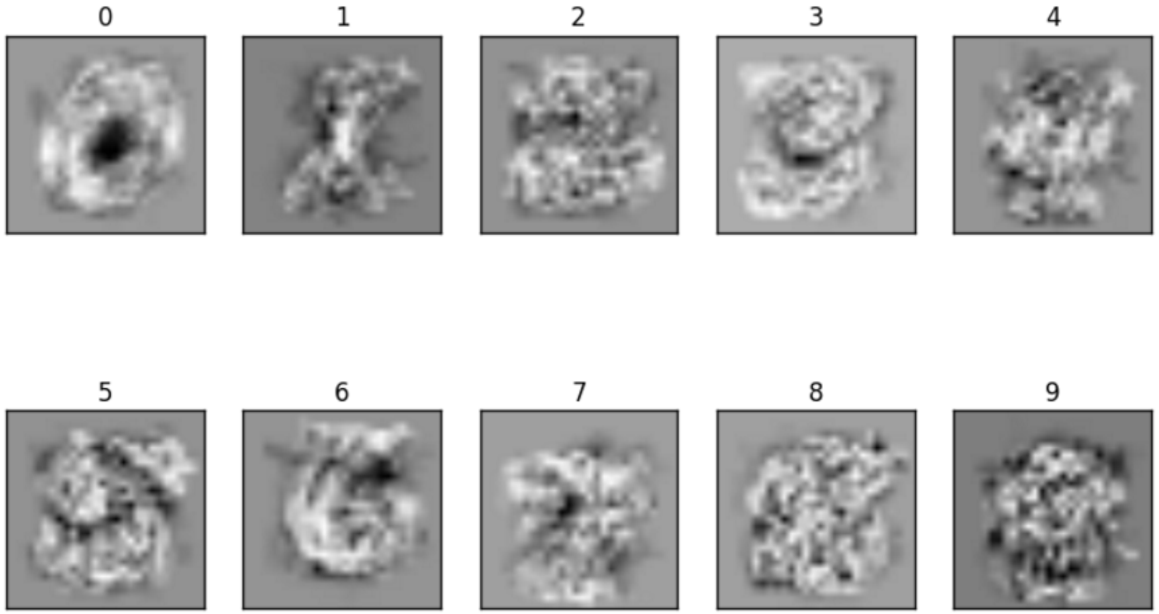
- Suppose $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$. Then: $z_k = \boldsymbol{w}_k^T \boldsymbol{x}$
  - $\boldsymbol{w}_k$ is 784-dim row of W
- When is z_k large?
- Theorem (proof on board): If $\boldsymbol{u}$ is any vector, then

$$\arg\max_{\|\boldsymbol{x}\|=1} \boldsymbol{u}^T \boldsymbol{x} = \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}$$

- Conclusion: For a given $\|\boldsymbol{x}\|$, $z_k = \boldsymbol{w}_k^T \boldsymbol{x}$ is maximized when $\boldsymbol{x} = \alpha \boldsymbol{w}_k$
  - Output of class k will be large when $\boldsymbol{x}$ is aligned with $\boldsymbol{w}_k$
  - Called the "matched filter" in signal processing

Visualizing the Weights

- Each class weight can be viewed as an image.
- Class weight output z_k will be large when it is aligned with $\boldsymbol{w}_k$



Optimized weights for logistic classifier

Why are they blurry?

Problems with Logistic Classifier

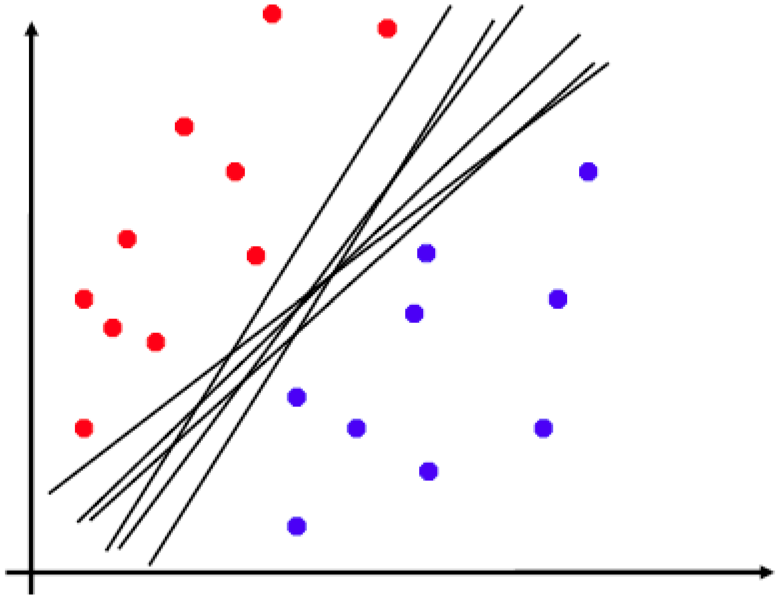
- Linear weighting cannot capture many deformities in image
  - Rotations
  - Translations
  - Variations in relative size of digit components
- Can be improved with preprocessing
  - E.g. deskewing, contrast normalization, many methods
- Is there a better classifier?

Outline

- Motivating example: Recognizing handwritten digits
  - Why logistic regression doesn't work well.
- Maximum margin classifiers
- Support vector machines
- Constrained optimization
- Kernel trick

Linear Separability and Non-Uniqueness of Separating plane

- When the samples are linearly separable, one can find a separating hyper-plane as a linear classifier.
- Separating hyper-plane is not unique
- Fig. on right: Many separating planes
- Which one is optimal?



Hyperplane Basics

- A hyperplane in d-dimensional space is defined by

$$b + w_1 x_1 + \cdots w_d x_d = 0 \;\; or \;\; b + \boldsymbol{w}^T \boldsymbol{x} = 0$$

- The parameters are unique only to a scaling factor:
  - $(b, \boldsymbol{w})$ and $(\alpha b, \alpha \boldsymbol{w})$ define the same plane.
  - For unique definition, we can require $\|\boldsymbol{w}\|=1$.
- The norm vector to the hyperplane is $\boldsymbol{w}/\|\boldsymbol{w}\|$.
- Distance of any point **x** to the hyperplane is $f(\boldsymbol{x})/\|\boldsymbol{w}\|$, where $f(\boldsymbol{x}) = b + \boldsymbol{w}^T \boldsymbol{x}$.
- See ESL Sec. 4.5.

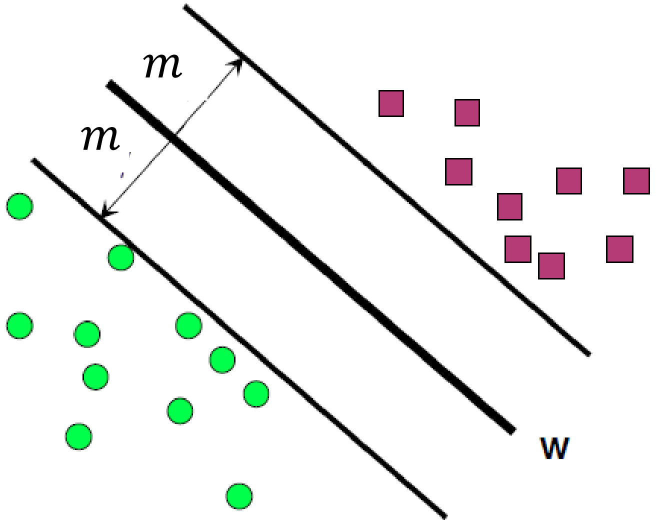
- ESL: Hastie, Tibshirani, Friedman, "The Elements of Statistical Learning". 2nd Ed. Springer.

Recap: Linear Separability and Margin

- Given training data $(\boldsymbol{x}_i, y_i), i = 1, \dots, N$
- Binary class label: $y_i = \pm 1$
- Perfectly linearly separable if there exists a $\boldsymbol{\theta} = (b, w_1, \dots, w_d)$ and $\gamma > 0$ s.t.:
  - $b + w_1 x_{i1} + \cdots w_d x_{id} > \gamma$ when $y_i = 1$
  - $b + w_1 x_{i1} + \cdots w_d x_{id} < -\gamma$ when $y_i = -1$
- $(\boldsymbol{w}, b)$ defines the separating hyperplane
- m is the margin: the minimal distance of a sample to the plane
- Single equation form:

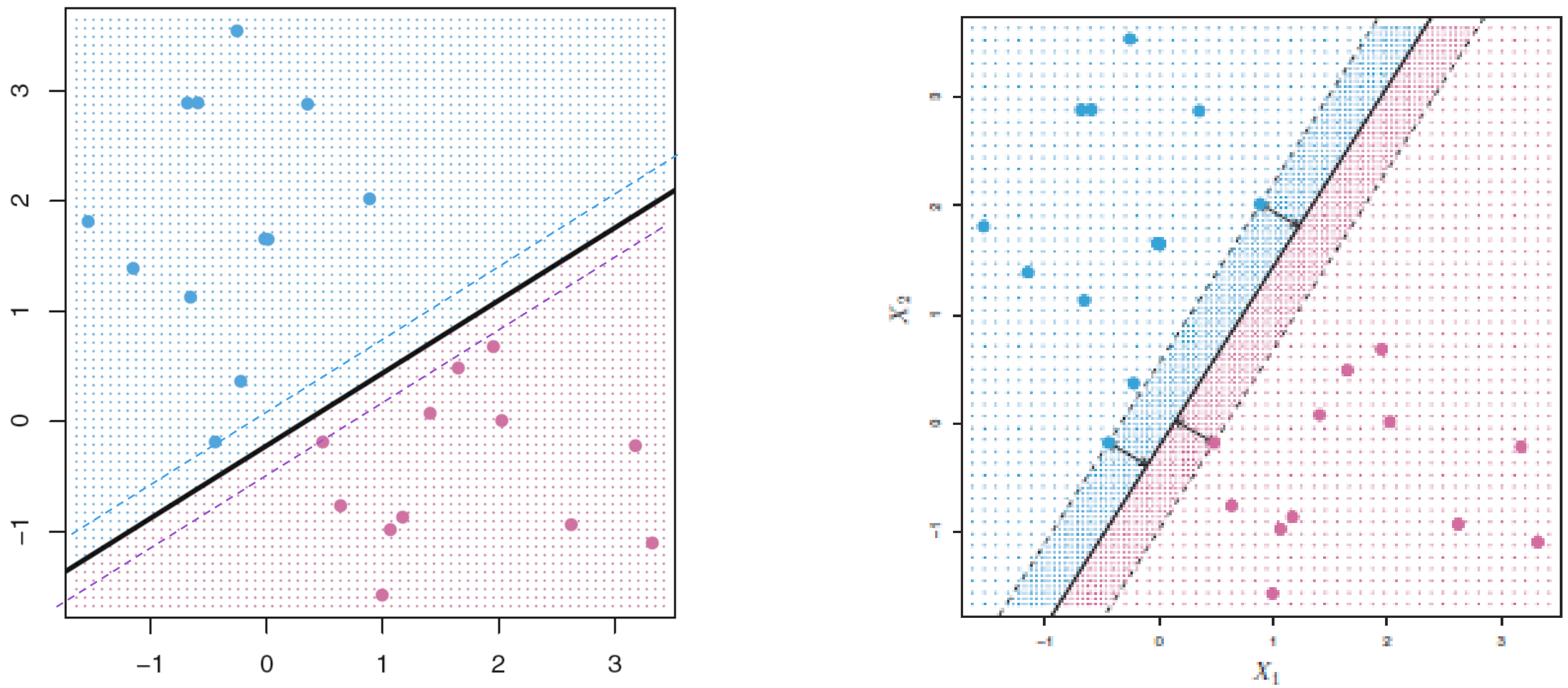
$$y_i(b + w_1 x_{i1} + \cdots w_d x_{id}) > \gamma \text{ for all } i = 1, \dots, N$$

$$\mathrm{m} = \frac{\gamma}{\|\boldsymbol{w}\|}$$



Recall that the distance of a point x to the line is $(b + w^T \boldsymbol{x})/\|\boldsymbol{w}\|$.
For points on the margin line, $b + w^T \boldsymbol{x} = \gamma$, distance m= $\gamma / \|\boldsymbol{w}\|$.

Which separating plane is better ?



From Fig. 9.2 and Fig. 9.3 in ISL.

Maximum Margin Classifier

- For the classifier to be more robust to noise, we want to maximize the margin!

- Define maximum margin classifier

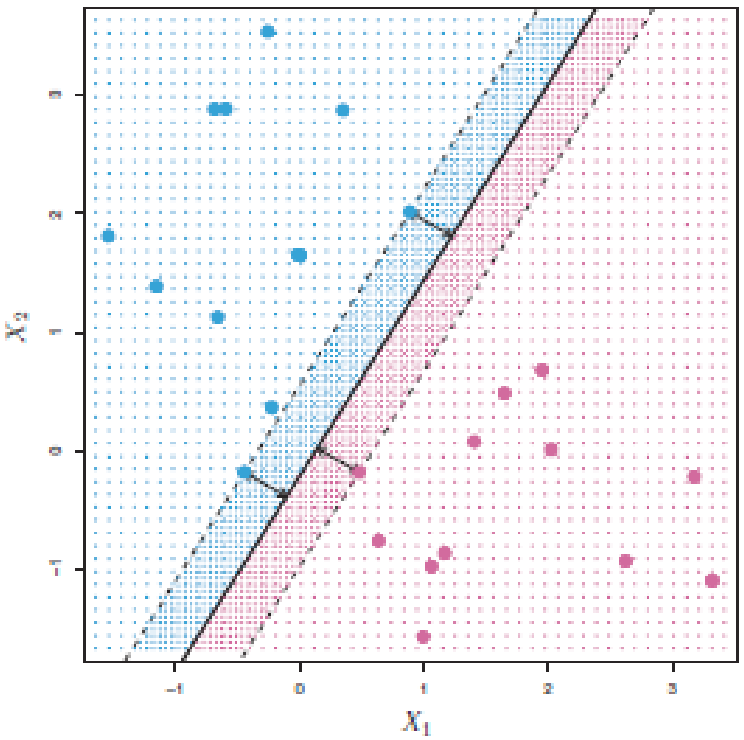
$$\max_{w,\gamma} \gamma$$ Maximizes the margin

  - Such that $y_i(b + \boldsymbol{w}^T \boldsymbol{x}) \geq \gamma \quad \text{for all } i$ Ensures all points are correctly classified
  - $\sum_{j=1}^{d} w_j^2 \leq 1$ Scaling on weights

- Called a constrained optimization
  - Objective function and constraints
  - More on this later.

- See closed form solution in Sec. 4.5.2 in ESL. Note notation difference.

Visualizing Maximum Margin Classifier



- ❑Fig. 9.3 of ISL
- ❑Margin determined by closest points to the line
  - The maximal margin hyperplane represents the mid-line of the widest “slab” that we can insert between two classes
- ❑In this figure, there are 3 points at the margin

ISL: James, Witten, Hastie, Tibshirani, An Introduction to Statistical Learning, Springer. 2013.

Problems with MM classifier

- Data is often not perfectly separable
  - Only want to correctly separate most points

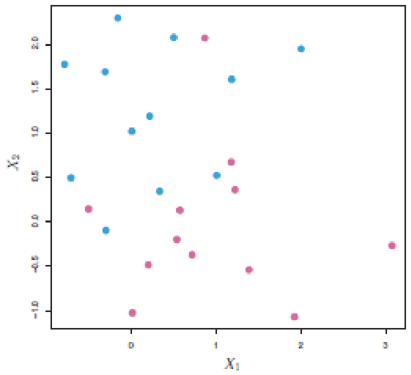


Fig. 9.4

- MM classifier is not robust
  - A single sample can radically change line

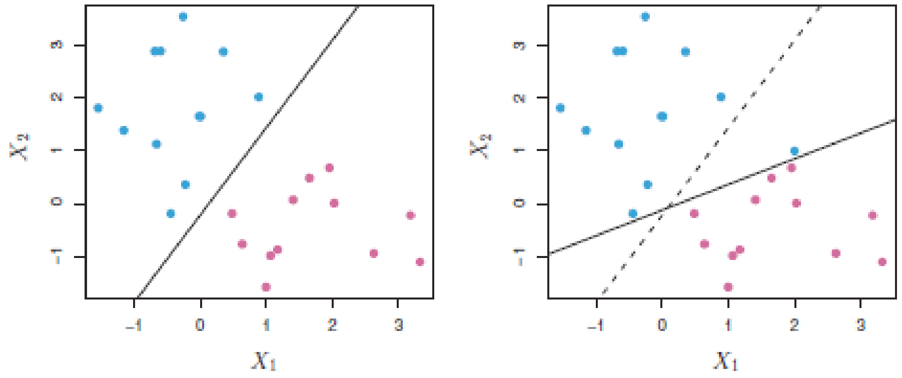


Fig. 9.5

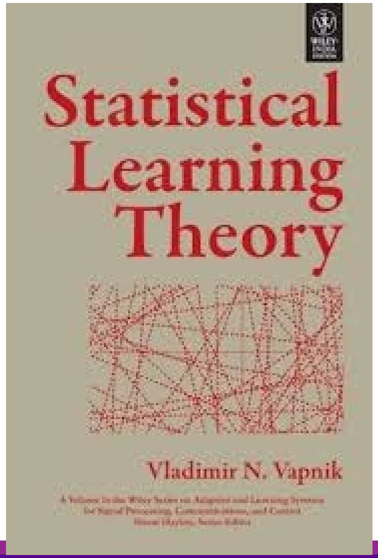
Outline

- Motivating example: Recognizing handwritten digits
  - Why logistic regression doesn't work well.
- Maximum margin classifiers
- Support vector machines
- Constrained optimization
- Kernel trick

Support Vector Machine

- Support Vector Machine (SVM)
  - Vladimir Vapnik, 1963
  - But became widely-used with kernel trick, 1993
  - More on this later

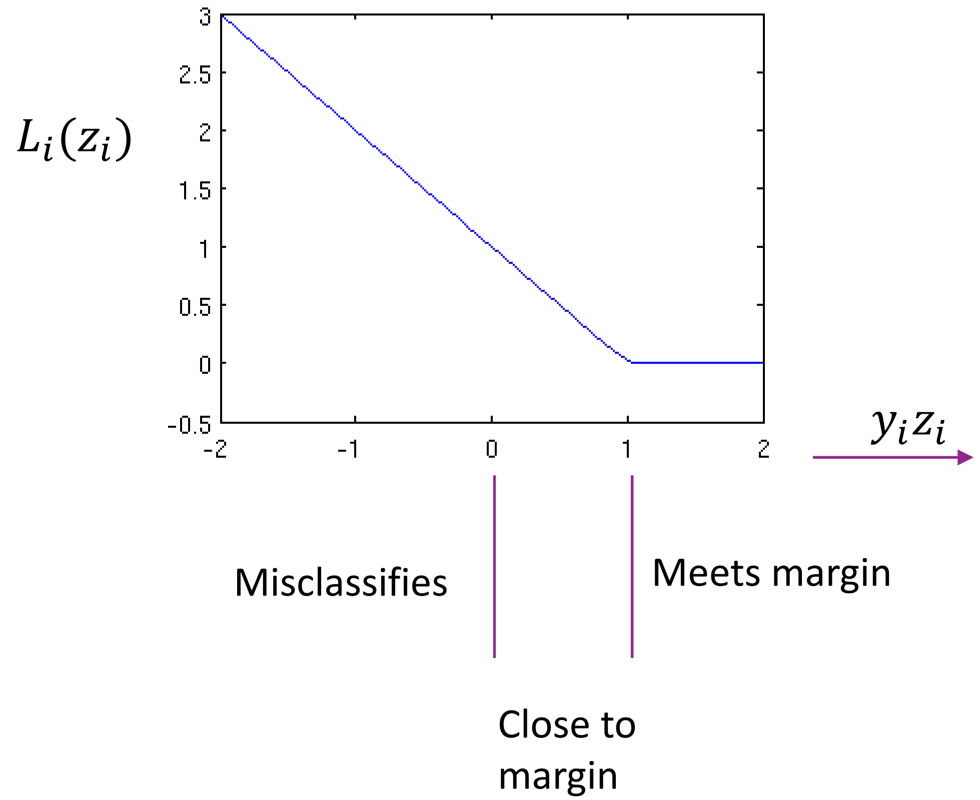
- Got best results on character recognition



- Key idea: Allow "slack" in the classification
  - Support vector classifier (SVC): Directly use raw features. Good when the original feature space is roughly linearly separable
  - Support vector machine (SVM): Map the raw features to some other domain through a kernel function

Hinge Loss

- Fix $\gamma = 1$
- Want ideally: $y_i(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1$ for all samples i
  - Equivalently, $y_i z_i \geq 1, \quad z_i = b + \boldsymbol{w}^T\boldsymbol{x}$
- But, perfect separation may not be possible
- Define hinge loss or soft margin:
  - $L_i(\mathbf{w}, b) = \max(0, 1 - y_i z_i)$
- Starts to increase as sample is misclassified:
  - $y_i z_i \geq 1 \;\Rightarrow$ Sample meets margin target, $L_i(w) = 0$
  - $y_i z_i \in [0,1) \;\Rightarrow$ Sample margin too small, small loss
  - $y_i z_i \leq 0 \;\Rightarrow$ Sample misclassified, large loss



SVM Optimization

- Given data $(\boldsymbol{x}_i, y_i)$
- Optimization $\min_{w,b} J(\boldsymbol{w}, b)$

$$J(\boldsymbol{w}, b) = C \sum_{i=1}^{N} \max(0, 1 - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)) + \frac{1}{2} ||\boldsymbol{w}||^2$$

Hinge loss term
Attempts to reduce
Misclassifications

C controls final margin

margin=$1/||\boldsymbol{w}||$

- Constant $C > 0$ will be discussed below
- Note: ISL book uses different naming conventions.
  - We have followed convention in sklearn

Alternate Form of SVM Optimization

- Equivalent optimization:

$$\min J_1(\boldsymbol{w}, b, \boldsymbol{\epsilon}), \qquad J_1(\boldsymbol{w}, b, \boldsymbol{\epsilon}) = C\sum_{i=1}^{N}\epsilon_i + \frac{1}{2}\|\boldsymbol{w}\|^2$$

- Subject to constraints:

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \epsilon_i \text{ for all } i = 1, \ldots, N$$

  - ϵ_i = amount sample i misses margin target

- Sometimes write as $J_1(\boldsymbol{w}, b, \boldsymbol{\epsilon}) = C\|\boldsymbol{\epsilon}\|_1 + \frac{1}{2}\|\boldsymbol{w}\|^2$

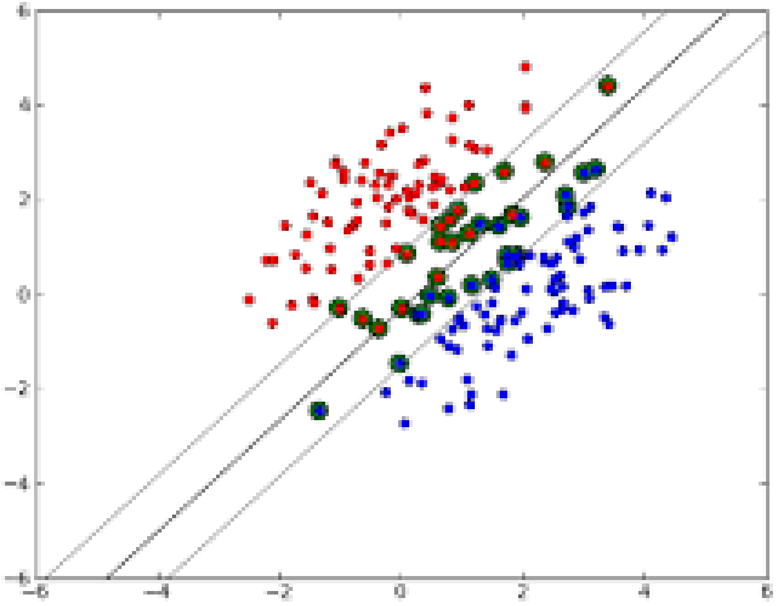
  - $\|\boldsymbol{\epsilon}\|_1 = \sum_{i=1}^{N}\epsilon_i$ called the "one-norm"
  - Generally one-norm would have absolute sign over ϵ_i. But in this case, when the constraint is met, ϵ_i>=0.

Interpreting Parameters

- Margin is $1/\|\boldsymbol{w}\|$
- Parameter ϵ_i called the slack variable
  - $\epsilon_i = 0 \Rightarrow$ Sample on correct side of margin
  - $0 \leq \epsilon_i < 1 \Rightarrow$ Sample violates the margin (are inside the margin)
  - $\epsilon_i \geq 1 \Rightarrow$ Sample misclassified (wrong side of hyperplane)
- Parameter C:
  - Balance between first term (violations) and second term (inverse of margin)
  - C large: Forces minimum number of violations, but small margin.
    - Highly fit to data. Low bias, higher variance
  - C small: Enables more samples violations, but large margin.
    - Higher bias, lower variance
  - Found by cross-validation

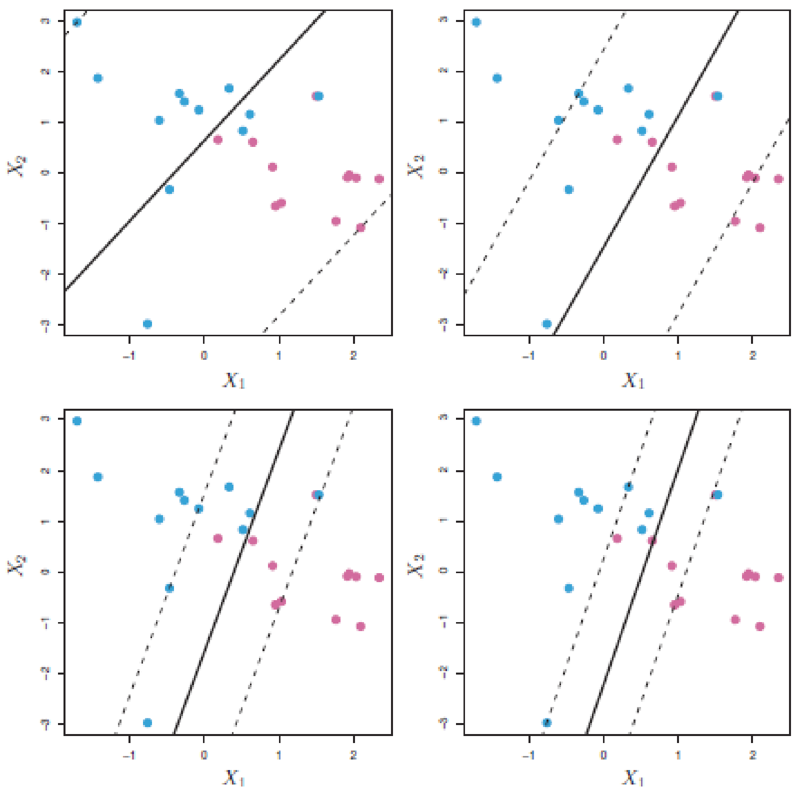
Support Vectors

- Support vectors: Samples that either:
  - Are exactly on margin: $y_{\mathrm{i}}(\boldsymbol{w}^T\boldsymbol{x}_i + b) = 1$
  - Or, on wrong side of margin: $y_{\mathrm{i}}(\boldsymbol{w}^T\boldsymbol{x}_i + b) \leq 1$
- Changing samples that are not SVs
  - Does not change solution
  - Provides robustness



Illustrating Effect of C

Low $C = C_1$ $\quad$ C_2



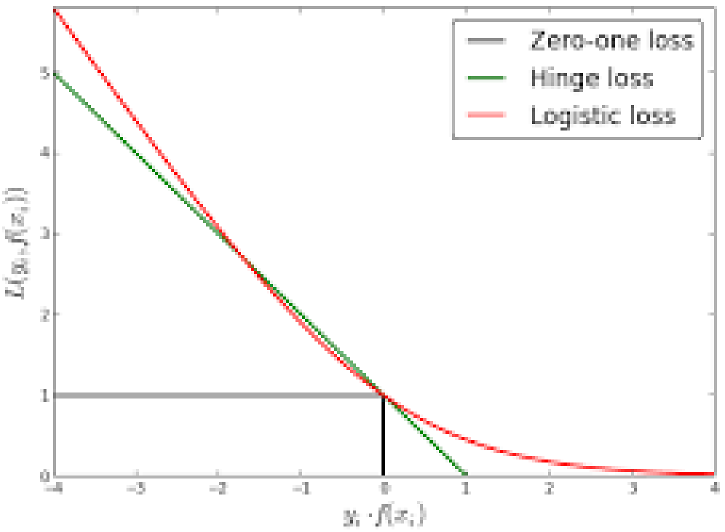
C_3 $\quad$ High C_4

- Fig. 9.7 of ISL
  - Note: C has opposite meaning in ISL than python
  - Here, we use python meaning
- Low C:
  - Leads to large margin
  - But allow many violations of margin.
  - Many more SVs
  - Reduces variance by using more samples
- Large C:
  - Leads to small margin
  - Reduce number of violations, and fewer SVs.
  - Highly fit to data. Low bias, higher variance
  - More chance to overfit

Relation to Logistic Regression

- Logistic regression also minimizes a loss function:

$$J(\boldsymbol{w}, b) = \sum_{i=1}^{N} L_i(w, b)\,, \qquad L_i(w, b) = \ln P(y_i|\boldsymbol{x}_i) = -\ln(1 + e^{-y_i z_i})$$



Outline

- ❑ Motivating example: Recognizing handwritten digits
  - Why logistic regression doesn't work well.
- ❑ Maximum margin classifiers
- ❑ Support vector machines
- ❑ Constrained optimization
- ❑ Kernel trick

Constrained Optimization

- In many problems, variables are constrained
- Constrained optimization formulation:
  - Objective: Minimize $f(\boldsymbol{w})$
  - Constraints: $g_1(\boldsymbol{w}) \leq 0, \ldots, g_M(\boldsymbol{w}) \leq 0$
- Examples:
  - Minimize the mpg of a car subject to a cost or meeting some performance
  - In ML: weight vector may have constraints from physical knowledge
- Often write constraints in vector form: Write $g(\boldsymbol{w}) \leq 0$

$$g(\boldsymbol{w}) = [g_1(\boldsymbol{w}), \ldots, g_m(\boldsymbol{w})]^T$$

Lagrangian

- Constrained optimization: Min $f(\mathbf{w})$ s.t. $g(\mathbf{w}) \leq 0$
- Consider first a single constraint: $g(\mathbf{w})$ is a scalar
- Define Lagrangian: $L(\mathbf{w}, \lambda) = f(\mathbf{w}) + \lambda g(\mathbf{w})$
  - $\mathbf{w}$ is called the primal variable
  - λ is called the dual variable
- Dual minimization: Given a dual parameter λ, minimize

$$\hat{\mathbf{w}}(\lambda) = \arg\min_{\mathbf{w}} L(\mathbf{w}, \lambda), \qquad L^*(\lambda) = \min_{w} L(\mathbf{w}, \lambda)$$

  - Minimizes a weighted combination of objective and constraint.
  - Higher $\lambda \Rightarrow$ Weight constraint more (try to make $g(\mathbf{w})$ smaller)
  - Lower $\lambda \Rightarrow$ Weight objective more (try to make $f(\mathbf{w})$ smaller)

KKT Conditions

- Given objective $f(\boldsymbol{w})$ and constraint $g(\boldsymbol{w})$
- KKT Conditions: $\widehat{\boldsymbol{w}}, \hat{\lambda}$ satisfy:
  - $\widehat{\boldsymbol{w}}$ minimizes the Lagrangian: $\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} L(\boldsymbol{w}, \hat{\lambda})$
  - Either
    - $g(\widehat{\boldsymbol{w}}) = 0$ and $\hat{\lambda} \geq 0$ [active constraint]
    - $g(\widehat{\boldsymbol{w}}) < 0$ and $\hat{\lambda} = 0$ [inactive constraint]
- Theorem: Under some technical conditions,
  - if $\widehat{\boldsymbol{w}}, \hat{\lambda}$ are local mimima of the constrained optimization, they must satisfy KKT conditions

General Procedure for Single Constraint

- Suppose:
  - $\boldsymbol{w} = (w_1, \dots, w_d)^T$: d unknown primal variables
  - $g(\boldsymbol{w}) \le 0$: scalar constraint
- Case 1: Assume constraint is active:
  - Solve $\boldsymbol{w}$ and λ: $\partial L(\boldsymbol{w}, \lambda)/\partial w_j = 0$ and $g(\boldsymbol{w}) = 0$ (resulting from setting $\partial L(\boldsymbol{w}, \lambda)/\partial \lambda = 0$)
  - $d + 1$ unknowns and $d + 1$ equations
  - Verify that $\lambda \ge 0$
- Case 2: Assume constraint is inactive
  - Solve primal objective $\partial f(\boldsymbol{w})/\partial w_j = 0$ ignoring constraint
  - d unknowns and d equations
  - Verify that constraint is satisfied: $g(\boldsymbol{w}) \le 0$

KKT Conditions Illustrated

- Example 1: Constraint is "active"

$$\min_w w^2 \quad s.t.\; w + 1 \le 0$$

- Example 2: Constraint is "inactive"

$$\min_w w^2 \quad s.t.\; w - 1 \le 0$$

- Examples worked on board with illustration

Multiple Constraints

- Now consider constraint: $g(\boldsymbol{w}) = [g_1(\boldsymbol{w}), \dots, g_M(\boldsymbol{w})]^T \leq 0$.
- Lagrangian is:

$$L(\boldsymbol{w}, \boldsymbol{\lambda}) = f(\boldsymbol{w}) + \boldsymbol{\lambda}^T g(\boldsymbol{w}) = f(\boldsymbol{w}) + \sum_{m=1}^{M} \lambda_m \, g_m(\boldsymbol{w})$$

  - Weighted sum of all M constraints
  - $\boldsymbol{\lambda}$ is called the dual vector

- KKT conditions extend to:
  - $\widehat{\boldsymbol{w}}$ minimizes the Lagrangian: $\widehat{\boldsymbol{w}} = \arg\min_{\boldsymbol{w}} L(\boldsymbol{w}, \hat{\lambda})$
  - For each $m = 1, \dots, M$
    - $g_m(\widehat{\boldsymbol{w}}) = 0$ and $\hat{\lambda}_m \geq 0$ [active constraint]
    - $g_m(\widehat{\boldsymbol{w}}) < 0$ and $\hat{\lambda}_m = 0$ [inactive constraint]

SVM Constrained Optimization

- Recall: SVM constrained optimization

$$\min J_1(\boldsymbol{w}, b, \boldsymbol{\epsilon}), \qquad J_1(\boldsymbol{w}, b, \boldsymbol{\epsilon}) = C \sum_{i=1}^{N} \epsilon_i + \frac{1}{2}\|\boldsymbol{w}\|^2$$

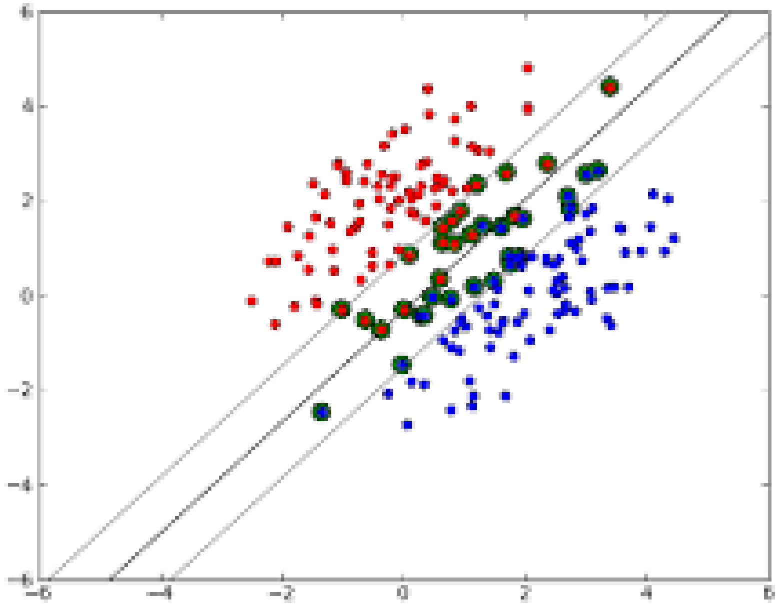
  - Constraints: $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 - \epsilon_i$ and $\epsilon_\mathrm{i} \geq 0$ for all $i = 1, \ldots, N$

- After applying KKT conditions and some algebra [beyond this class], solution is
  - Optimal weight vector: $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$ linear combination of instances
  - Dual parameters α_i minimize

$$\sum_{i=1}^{N} \alpha_i + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j \quad \mathrm{s.t.}\ \ 0 \leq \alpha_i \leq C$$

Support Vectors

- Classifier weight is: $\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$
- Can show that $\alpha_i > 0$ only when $\boldsymbol{x}_i$ is a support vector
  - On boundary or violating constraint
  - Otherwise $\alpha_i = 0$



Correlation Interpretation of SVM

- Classifier weight is:

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i$$

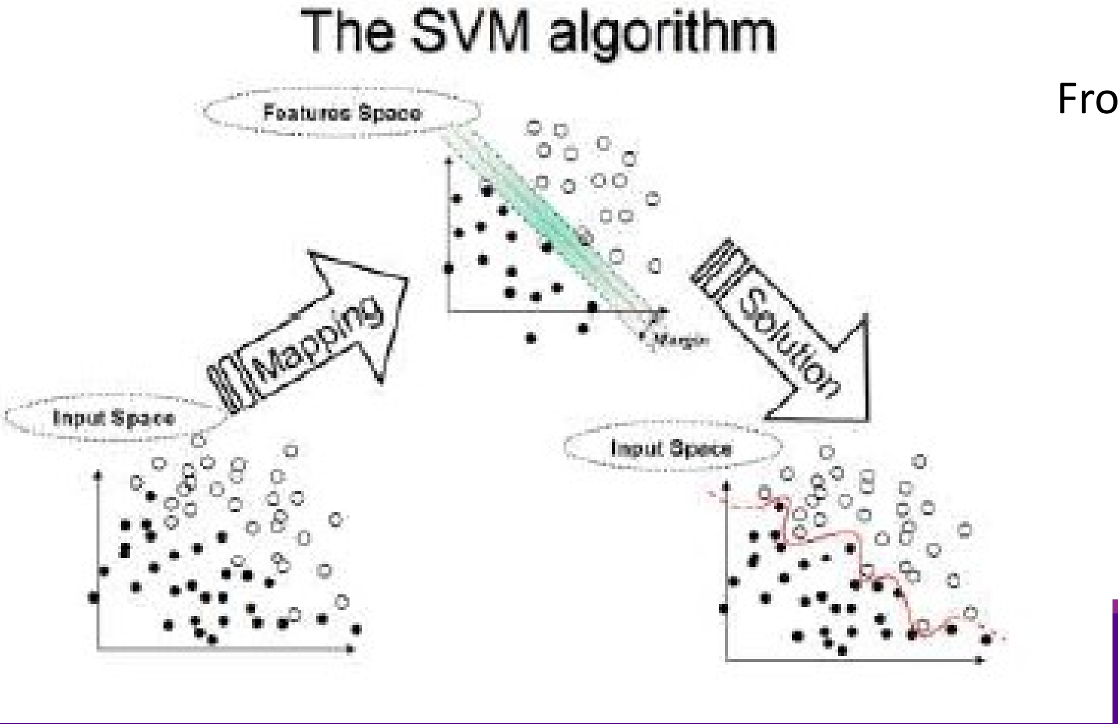
- Now suppose we are given a new sample $\boldsymbol{x}$ to classify
- Classifier discriminant function for any test sample $\boldsymbol{x}$ is:
  - $\hat{z}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b = \sum \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + b$
- Classifier output
  - $\hat{y}(\boldsymbol{x}) = \text{sign}(\hat{z}(\boldsymbol{x}))$
  - Measure "correlation" $\boldsymbol{x}_i^T \boldsymbol{x}$ of new sample $\boldsymbol{x}$ with each support vector $\boldsymbol{x}_i$ in training data
  - Predicted label depends on the weighted average of labels for the support vectors, with weights proportional to the correlation of the test sample with the support vector.

Outline

- Motivating example: Recognizing handwritten digits
  - Why logistic regression doesn't work well.
- Maximum margin classifiers
- Support vector machines
- Constrained optimization
- Kernel trick

Transform Problem

- Transform problem: replace $\boldsymbol{x}$ with $\phi(\boldsymbol{x})$
  - Enables more rich, non-linear classifiers
  - Examples: polynomial classification $\phi(x) = [1, x, x^2, \ldots, x^{d-1}]$
- Tries to find separation in a feature space



From https://www.dtreg.com/solution/view/20

Transform Problem

- SVM problem in transformed domain:

$$J(\boldsymbol{w}, b) = C \sum_{i=1}^{N} \max(0, 1 - y_i(\boldsymbol{w}^T \phi(\boldsymbol{x}_i) + b)) + \frac{1}{2} \|\boldsymbol{w}\|^2$$

- Solution is of the form:

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \phi(\boldsymbol{x}_i)$$

- Classifier discriminant function:

$$z = b + \boldsymbol{w}^T \boldsymbol{x} = b + \sum_{i=1}^{N} \alpha_i y_i \underbrace{\phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x})}_{K(\boldsymbol{x}_i, \boldsymbol{x}) = \text{“kernel”}}$$

Kernel Trick

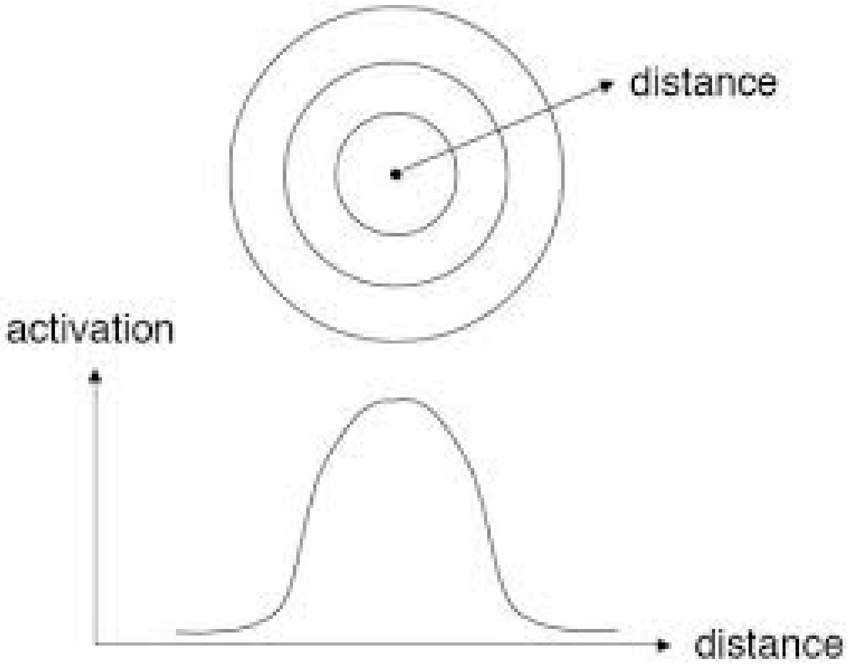
- Classifier is:
  - $z = b + \boldsymbol{w}^T\boldsymbol{x} = b + \sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x})$, $\quad K(\boldsymbol{x}_i, \boldsymbol{x}) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x})$
  - $\hat{y} = \text{sign}(z) = \begin{cases} 1 & \text{if } z > 0 \\ -1 & \text{if } z < 0 \end{cases}$

- Do not need to explicitly compute $\phi(\boldsymbol{x})$
- Can directly compute kernel $K(\boldsymbol{x}_i, \boldsymbol{x})$
  - Provided kernel corresponds to some $\phi(\boldsymbol{x})$

Understanding the Kernel

- Kernel function $K(\boldsymbol{x}_i, \boldsymbol{x})$:
  - measures "similarity" between new sample $\boldsymbol{x}$ and training data $\boldsymbol{x}_i$
  - $K(\boldsymbol{x}_i, \boldsymbol{x})$ large $\Rightarrow \boldsymbol{x}_i, \boldsymbol{x}$ close
  - $K(\boldsymbol{x}_i, \boldsymbol{x}) \approx 0 \Rightarrow \boldsymbol{x}_i, \boldsymbol{x}$ far

- Linear discriminant $z = b + \sum_{i=1}^{N} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x})$
  - Weighs sample $\boldsymbol{x}_i$ that are close to $\boldsymbol{x}$

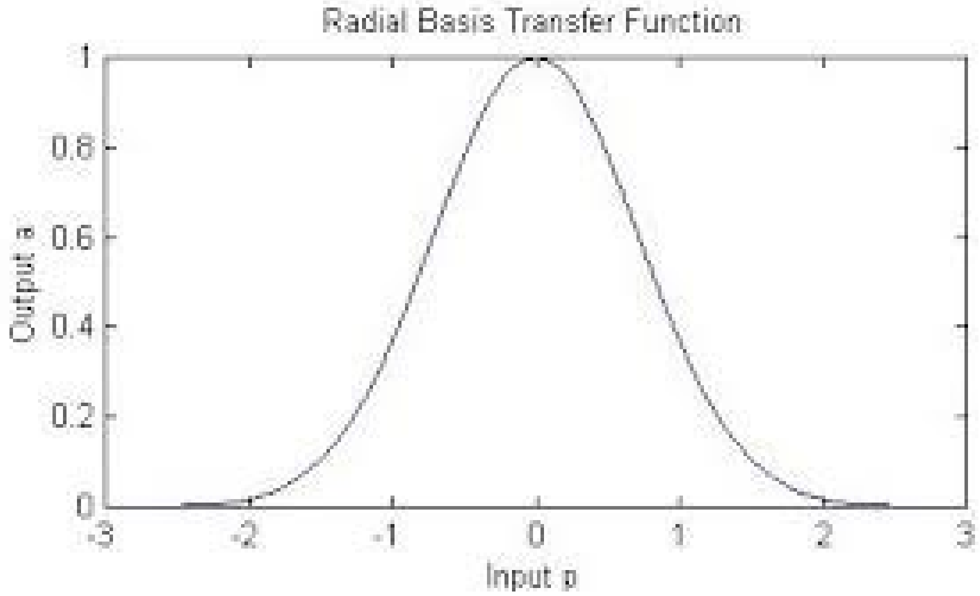


Possible Kernels

- Radial basis function:

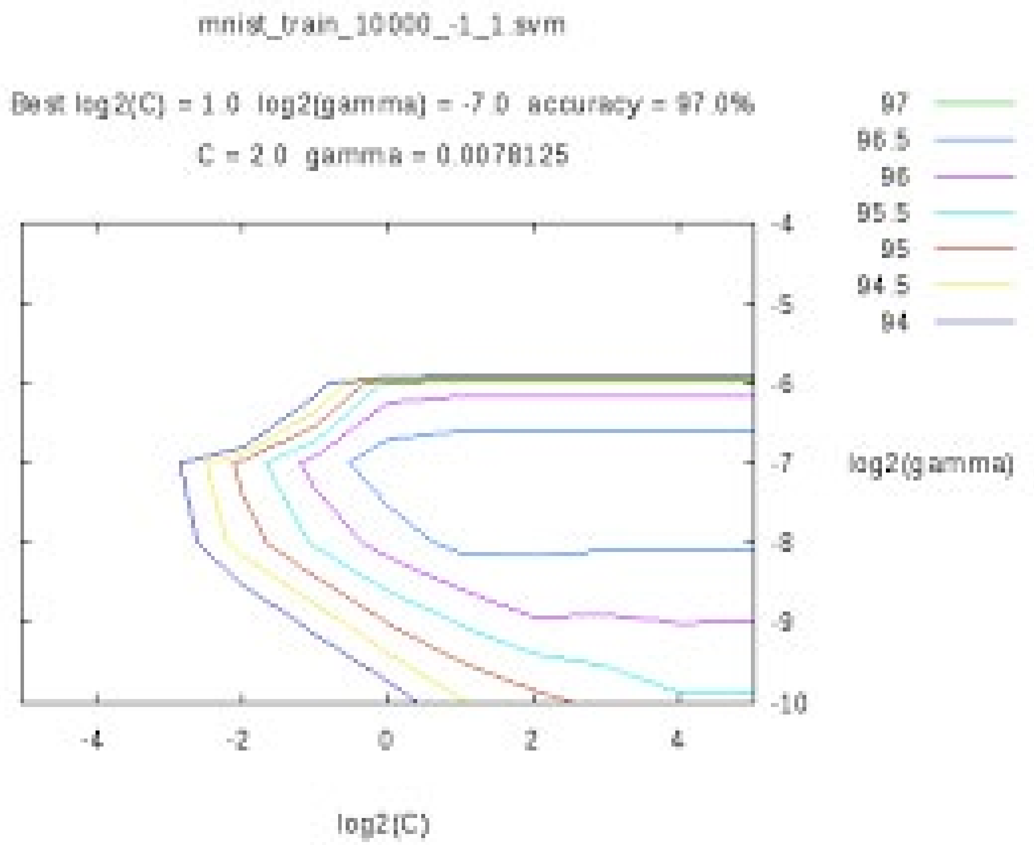
$$K(x, x') = \exp[-\gamma \|x - x'\|^2]$$

  - $1/\gamma$ indicates width of kernel
- Polynomial kernel: $K(x, x') = |x^T x|^d$
  - Typically d=2



Parameter Selection

- Consider SVM with:
  - Parameter $C > 0$, RBF with $\gamma > 0$
- Higher C or γ
  - Fewer SVs
  - Classifiers averages over smaller set
  - Lower bias, but higher variance
- Typically select via cross-validation
  - Try out different (C, γ)
  - Find which one provides highest accuracy on test set
- Python can automatically do grid search



http://peekaboo-vision.blogspot.com/2010/09/mnist-for-ever.html

Multi-Class SVMs

- Suppose there are K classes
- One-vs-one:
  - Train $\binom{K}{2}$ SVMs for each pair of classes
  - Test sample assigned to class that wins "majority of votes"
  - Best results but very slow
- One-vs-rest:
  - Train K SVMs: train each class k against all other classes
  - Pick class with highest z_k
- Sklearn has both options

MNIST Results

- ❑ Run classifier
- ❑ Very slow
  - Several minutes for 40,000 samples
  - Slow in training and test
  - Major drawback of SVM
- ❑ Accuracy ≈ 0.984
  - Much better than logistic regression
- ❑ Can get better with:
  - pre-processing
  - More training data
  - Optimal parameter selection

```
from sklearn import svm

# Create a classifier: a support vector classifier
svc = svm.SVC(probability=False,  kernel="rbf", C=2.8, gamma=.0073,verbose=10)
```

```
svc.fit(Xtr,ytr)
```

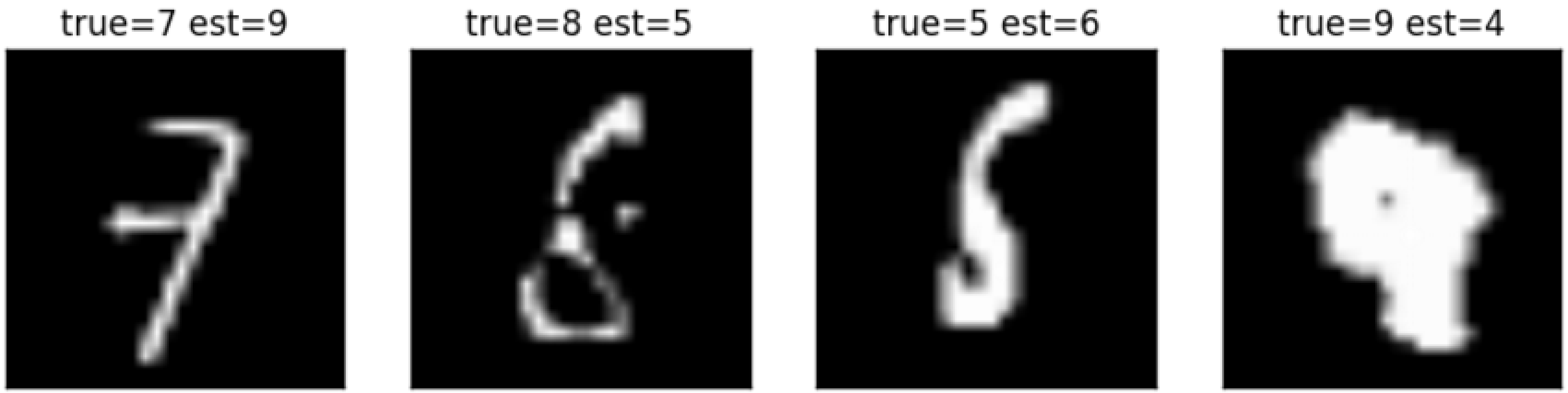
```
[LibSVM]

SVC(C=2.8, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape=None, degree=3, gamma=0.0073, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=10)
```

```
yhat1 = svc.predict(Xts)
acc = np.mean(yhat1 == yts)
print('Accuaracy = {0:f}'.format(acc))
```

```
Accuaracy = 0.984000
```

MNIST Errors

- Some of the error are hard even for a human



What you should know

- ❑Interpret weights in linear classification of images (logistic regression): Match filters
- ❑Understand the margin in linear classification and maximum margin classifier
- ❑SVM classifier: Allow violation of margin by introducing slack variables (More robust than linear classifier)
- ❑Solve constrained optimization using the Lagrangian.
  - Understand KKT conditions for a single constraint
- ❑Extend to nonlinear classifier by feature transformation: SVM with nonlinear kernels
- ❑Select SVM parameters from cross-validation