This set of slides gives a real example of using dual problems

- Basic concepts: SVM and kernels
- SVM primal/dual problems
- Logistic Regression
- Loss Functions
Outline

- Basic concepts: SVM and kernels
- SVM primal/dual problems
- Logistic Regression
- Loss Functions
Data Classification

- Given training data in different classes (labels known)
- Predict test data (labels unknown)
- Training and testing
Support Vector Classification

- **Training vectors**: \( x_i, i = 1, \ldots, l \)
- Feature vectors. For example,
  A patient = \([\text{height, weight, } \ldots]\)^T
- Consider a simple case with two classes:
  Define an *indicator* vector \( y \)
  \[
  y_i = \begin{cases} 
  1 & \text{if } x_i \text{ in class 1} \\
  -1 & \text{if } x_i \text{ in class 2}
  \end{cases}
  \]
- A hyperplane which separates all data
A separating hyperplane: $\mathbf{w}^T \mathbf{x} + b = 0$

$$(\mathbf{w}^T \mathbf{x}_i) + b \geq 1 \quad \text{if } y_i = 1$$

$$(\mathbf{w}^T \mathbf{x}_i) + b \leq -1 \quad \text{if } y_i = -1$$

Decision function $f(\mathbf{x}) = \text{sgn}(\mathbf{w}^T \mathbf{x} + b)$, $\mathbf{x}$: test data

Many possible choices of $\mathbf{w}$ and $b$
Maximal Margin

- Distance between $\mathbf{w}^T \mathbf{x} + b = 1$ and $-1$:
  
  $$\frac{2}{\|\mathbf{w}\|} = \frac{2}{\sqrt{\mathbf{w}^T \mathbf{w}}}$$

- A quadratic programming problem (Boser et al., 1992)

$$\begin{align*}
\min_{\mathbf{w}, b} & \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} \\
\text{subject to} & \quad y_i (\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \\
\end{align*}$$

\[i = 1, \ldots, l.\]
Data May Not Be Linearly Separable

- An example:

\[ \phi(x) = [\phi_1(x), \phi_2(x), \ldots]^T. \]
Standard SVM (Boser et al., 1992; Cortes and Vapnik, 1995)

\[
\begin{align*}
\min_{\mathbf{w}, b, \xi} & \quad \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^{l} \xi_i \\
\text{subject to} & \quad y_i (\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\
& \quad \xi_i \geq 0, \quad i = 1, \ldots, l.
\end{align*}
\]

Example: \( \mathbf{x} \in \mathbb{R}^3, \phi(\mathbf{x}) \in \mathbb{R}^{10} \)

\[
\phi(\mathbf{x}) = \begin{bmatrix}
1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_3, x_1^2, \\
x_2^2, x_3^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, \sqrt{2}x_2x_3
\end{bmatrix}^T
\]
Finding the Decision Function

- \( \mathbf{w} \): \textbf{maybe infinite} variables
- The \textbf{dual} problem: \textbf{finite} number of variables

\[
\begin{align*}
\min_{\alpha} & \quad \frac{1}{2} \alpha^T Q \alpha - \mathbf{e}^T \alpha \\
\text{subject to} & \quad 0 \leq \alpha_i \leq C, \ i = 1, \ldots, l \\
& \quad \mathbf{y}^T \alpha = 0,
\end{align*}
\]

where \( Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j) \) and \( \mathbf{e} = [1, \ldots, 1]^T \)

- At optimum

\[
\mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)
\]

- A \textbf{finite} problem: \#variables = \#training data
Kernel Tricks

- $Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j)$ needs a **closed** form
- Example: $x_i \in \mathbb{R}^3$, $\phi(x_i) \in \mathbb{R}^{10}$

$$\phi(x_i) = [1, \sqrt{2}(x_i)_1, \sqrt{2}(x_i)_2, \sqrt{2}(x_i)_3, (x_i)_1^2, (x_i)_2^2, (x_i)_3^2, \sqrt{2}(x_i)_1(x_i)_2, \sqrt{2}(x_i)_1(x_i)_3, \sqrt{2}(x_i)_2(x_i)_3]^T$$

Then $\phi(x_i)^T \phi(x_j) = (1 + x_i^T x_j)^2$.

- Kernel: $K(x, y) = \phi(x)^T \phi(y)$; common kernels:
  - $e^{-\gamma \|x_i - x_j\|^2}$, (Radial Basis Function)
  - $(x_i^T x_j / a + b)^d$ (Polynomial kernel)
Can be inner product in infinite dimensional space

Assume $x \in R^1$ and $\gamma > 0$.

$$e^{-\gamma \|x_i-x_j\|^2} = e^{-\gamma (x_i-x_j)^2} = e^{-\gamma x_i^2 + 2\gamma x_i x_j - \gamma x_j^2}$$

$$= e^{-\gamma x_i^2 - \gamma x_j^2} \left( 1 + \frac{2\gamma x_i x_j}{1!} + \frac{(2\gamma x_i x_j)^2}{2!} + \frac{(2\gamma x_i x_j)^3}{3!} + \cdots \right)$$

$$= e^{-\gamma x_i^2 - \gamma x_j^2} \left( 1 \cdot 1 + \sqrt{\frac{2\gamma}{1!}} x_i \cdot \sqrt{\frac{2\gamma}{1!}} x_j + \sqrt{\frac{(2\gamma)^2}{2!}} x_i^2 \cdot \sqrt{\frac{(2\gamma)^2}{2!}} x_j^2 \right.$$

$$+ \sqrt{\frac{(2\gamma)^3}{3!}} x_i^3 \cdot \sqrt{\frac{(2\gamma)^3}{3!}} x_j^3 + \cdots \big) = \phi(x_i)^T \phi(x_j),$$

where

$$\phi(x) = e^{-\gamma x^2} \left[ 1, \sqrt{\frac{2\gamma}{1!}} x, \sqrt{\frac{(2\gamma)^2}{2!}} x^2, \sqrt{\frac{(2\gamma)^3}{3!}} x^3, \cdots \right]^T.$$
Basic concepts: SVM and kernels

**Decision function**

- At optimum

\[ \mathbf{w} = \sum_{i=1}^{l} \alpha_i y_i \phi(x_i) \]

- Decision function

\[
\begin{align*}
\mathbf{w}^T \phi(x) + b \\
= \sum_{i=1}^{l} \alpha_i y_i \phi(x_i)^T \phi(x) + b \\
= \sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b
\end{align*}
\]

- Only \( \phi(x_i) \) of \( \alpha_i > 0 \) used \( \Rightarrow \) support vectors
Support Vectors: More Important Data

Only $\phi(x_i)$ of $\alpha_i > 0$ used $\Rightarrow$ support vectors
Outline

- Basic concepts: SVM and kernels
- SVM primal/dual problems
- Logistic Regression
- Loss Functions
Deriving the Dual

- Consider the problem without $\xi_i$

$$\begin{align*}
\min_{w,b} & \quad \frac{1}{2} w^T w \\
\text{subject to} & \quad y_i(w^T \phi(x_i) + b) \geq 1, \ i = 1, \ldots, l.
\end{align*}$$

- Its dual

$$\begin{align*}
\min_{\alpha} & \quad \frac{1}{2} \alpha^T Q \alpha - e^T \alpha \\
\text{subject to} & \quad 0 \leq \alpha_i, \quad i = 1, \ldots, l, \\
y^T \alpha = 0.
\end{align*}$$
Lagrangian Dual

\[
\max_{\alpha \geq 0} \left( \min_{w, b} L(w, b, \alpha) \right),
\]

where

\[
L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{l} \alpha_i \left( y_i (w^T \phi(x_i) + b) - 1 \right)
\]

Strong duality

\[
\min \text{ Primal} = \max_{\alpha \geq 0} \left( \min_{w, b} L(w, b, \alpha) \right)
\]
Simplify the dual. When $\alpha$ is fixed,

$$
\min_{w, b} L(w, b, \alpha) =
\begin{cases}
-\infty & \text{if } \sum_{i=1}^{l} \alpha_i y_i \neq 0 \\
\min_w \frac{1}{2} w^T w - \sum_{i=1}^{l} \alpha_i [y_i (w^T \phi(x_i) - 1)] & \text{if } \sum_{i=1}^{l} \alpha_i y_i = 0
\end{cases}
$$

If $\sum_{i=1}^{l} \alpha_i y_i \neq 0$,

decrease

$$
-b \sum_{i=1}^{l} \alpha_i y_i
$$

in $L(w, b, \alpha)$ to $-\infty$
If $\sum_{i=1}^{l} \alpha_i y_i = 0$, optimum of the strictly convex

$$\frac{1}{2} w^T w - \sum_{i=1}^{l} \alpha_i [y_i (w^T \phi(x_i) - 1]$$

happens when

$$\frac{\partial}{\partial w} L(w, b, \alpha) = 0.$$

Thus,

$$w = \sum_{i=1}^{l} \alpha_i y_i \phi(x_i).$$
Note that

$$\mathbf{w}^T \mathbf{w} = \left( \sum_{i=1}^{l} \alpha_i y_i \phi(\mathbf{x}_i) \right)^T \left( \sum_{j=1}^{l} \alpha_j y_j \phi(\mathbf{x}_j) \right)$$

$$= \sum_{i,j} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

The dual is

$$\max_{\alpha \geq 0} \left\{ \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j) \right\}$$

if $$\sum_{i=1}^{l} \alpha_i y_i = 0$$,

if $$\sum_{i=1}^{l} \alpha_i y_i \neq 0$$. 
Lagrangian dual: \( \max_{\alpha \geq 0} \left( \min_{w, b} L(w, b, \alpha) \right) \)

\(-\infty \) definitely not maximum of the dual

Dual optimal solution not happen when

\[
\sum_{i=1}^{l} \alpha_i y_i \neq 0
\]

Dual simplified to

\[
\max_{\alpha \in \mathbb{R}^l} \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \phi(x_i)^T \phi(x_j)
\]

subject to \( y^T \alpha = 0 \),

\( \alpha_i \geq 0 \), \( i = 1, \ldots, l \).
Our problems may be infinite dimensional
Can still use Lagrangian duality
See a rigorous discussion in Lin (2001)
Outline

- Basic concepts: SVM and kernels
- SVM primal/dual problems
- Logistic Regression
- Loss Functions
For a label-feature pair \((y,x)\), assume the probability model

\[
p(y|x) = \frac{1}{1 + e^{-yw^T x}}.
\]

\(w\) is the parameter to be decided

Assume

\[(y_i, x_i), i = 1, \ldots, l\]

are training instances
Logistic regression finds $w$ by maximizing the following likelihood

$$\max_w \prod_{i=1}^{l} p(y_i|x_i).$$  \hfill (1)\hfill

Regularized logistic regression

$$\min_w \frac{1}{2} w^T w + C \sum_{i=1}^{l} \log \left( 1 + e^{-y_i w^T x_i} \right).$$  \hfill (2)\hfill

$C$: regularization parameter decided by users.
Outline

- Basic concepts: SVM and kernels
- SVM primal/dual problems
- Logistic Regression
- Loss Functions
- We derive SVM from the viewpoint of maximal margin.
- We derive logistic regression from minimizing the negative log likelihood.
- They can both be considered from the viewpoint of regularized linear classification.
Regularized Linear Classification

- Training data \( \{y_i, x_i\}, x_i \in R^n, i = 1, \ldots, l, y_i = \pm 1 \)
- \( l \): # of data, \( n \): # of features

\[
\min_w f(w), \quad f(w) \equiv \frac{w^T w}{2} + C \sum_{i=1}^{l} \xi(w; x_i, y_i)
\]

- \( w^T w / 2 \): regularization term (we have no time to talk about L1 regularization here)
- \( \xi(w; x, y) \): loss function: we hope \( yw^T x > 0 \)
- \( C \): regularization parameter
Some commonly used ones:

\[ \xi_{L1}(w; x, y) \equiv \max(0, 1 - yw^Tx), \quad (3) \]

\[ \xi_{L2}(w; x, y) \equiv \max(0, 1 - yw^Tx)^2, \quad (4) \]

\[ \xi_{LR}(w; x, y) \equiv \log(1 + e^{-yw^Tx}). \quad (5) \]

- SVM (Boser et al., 1992; Cortes and Vapnik, 1995): (3)-(4)
- Logistic regression (LR): (5)
Their performance is usually similar
However,

\[ \xi_{L1}: \text{not differentiable} \]
\[ \xi_{L2}: \text{differentiable but not twice differentiable} \]
\[ \xi_{LR}: \text{twice differentiable} \]

The same optimization method may not be applicable to all these losses
You can use $\|w\|_1$ regularization. This is now popular because of sparsity (i.e., some w’s components are zeros)

But do we still have maximal margin interpretation?

For SVM, can we have an interpretation like maximum likelihood of logistic regression?

For regularized logistic regression, can we have an interpretation of maximal margin?
