Announcements

- PS Nectarine due today; PS
Announcements

- PS Nectarine due today; PS ????? out by Monday.
  - Shorter than usual!

- 10/13: Guest lecture: Meinolf Sellman
  - Optimization, Lagrange multipliers, . . .

- 10/16: no class.

- 10/18: Guest lecture: Chad Jenkins
  - Robot learning, intro to unsupervised and reinforcement learning.
Review

- **Ridge regression:**

  \[
  \hat{w}_{\text{ridge}} = \arg\max_w - \sum_{i=1}^{N} (y_i - w^T x_i)^2 - \lambda \sum_{j=1}^{d} w_j^2 = (\lambda I + X^T X)^{-1} X^T y.
  \]

- **Lasso:**

  \[
  \hat{w}_{\text{lasso}} = \arg\max_w - \sum_{i=1}^{N} (y_i - w^T x_i)^2 - \lambda \sum_{j=1}^{d} |w_j|,
  \]

  \[
  \Rightarrow \text{no closed-form solution; must solve optimization problem}
  \]

  \[
  \arg\max_w \quad - \sum_{i=1}^{N} (y_i - w^T x_i)^2.
  \]

  \[
  \text{w: } \sum_{j=1}^{d} |w_j| \leq \beta
  \]

- **Choice of } \lambda \text{ or } \beta: \text{ by cross-validation.}
Plan for today

- Max-margin classification
- Support Vector Machines
  - focus on linearly separable case
The classification margin

Since the data are separable, we can find \( \mathbf{w} \) such that

\[
\forall i = 1, \ldots, N \quad y_i (w_0 + \mathbf{w}^T \mathbf{x}_i) > 0.
\]

We can even guarantee (by increasing \( \|\mathbf{w}\| \) if necessary)

\[
\forall i = 1, \ldots, N \quad y_i (w_0 + \mathbf{w}^T \mathbf{x}_i) \geq 1.
\]

\[
\min_i y_i (w_0 + \mathbf{w}^T \mathbf{x}_i) \text{ is the smallest distance from } \mathbf{x}_i \text{ to the boundary (half the separation between classes).}
\]

We will refer to it as the margin.
Max-margin boundary

- If we just state that we want $$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \min_i y_i(w_0 + \mathbf{w}^T \mathbf{x}_i)?$$

we run into the same problem we have seen with LR: when data are separable the margin is unbounded as $$\|\mathbf{w}\| \rightarrow \infty$$.

- Suppose $$y = 1$$, and $$\|\mathbf{w}\| = 1$$. Let $$w_0 + \mathbf{w}^T \mathbf{x} = c$$. Then,

$$\alpha w_0 + (\alpha \cdot \mathbf{w})^T \mathbf{w} = \alpha (w_0 + \mathbf{w}^T \mathbf{x}) = \alpha c,$$

i.e. we can achieve arbitrarily wide margin with the same classification boundary.

- One solution: require $$\|\mathbf{w}\| = 1$$. 


Fixed margin solution

- A more convenient solution: require fixed margin of, say, 1.
- Of all $w$ that achieve such margin, choose the smallest one.
  - This imposes a unique (equivalent) solution!
- The margin constraints, graphically (in 1D):

$$1 \cdot (w_0 + w_1 x_i) - 1 \geq 0, \quad y_i = 1$$
$$-1 \cdot (w_0 + w_1 x_i) - 1 \geq 0, \quad y_i = -1.$$
Fixed margin solution

- A more convenient solution: require fixed margin of, say, 1.
- Of all \( w \) that achieve such margin, choose the smallest one.
  - This imposes a unique (equivalent) solution!
- The margin constraints, graphically (in 1D):

\[
1 \cdot (w_0 + w_1 x_i) - 1 \geq 0, \quad y_i = 1
\]
\[
-1 \cdot (w_0 + w_1 x_i) - 1 \geq 0, \quad y_i = -1.
\]
Separation is maximal when the line passes through \((x^+ + x^-)/2\).
- The maximum margin is 1;
- the margin is \textit{inversely proportional} to the slope \(|w_1|\);
- The optimal boundary is achieved with margin \(1/|w_1|\),

\[
|w_1| = 2/|x^+ - x^-|.
\]
Margin and regularization

• In general $d$-dimensional case, we solve the regularization problem:

$$\text{minimize} \quad \|\mathbf{w}\|^2 = \sum_{j=1}^{d} w_j^2,$$

subject to the margin constraint

$$\forall i, \quad y_i(w_0 + \mathbf{w}^T \mathbf{x}_i) - 1 \geq 0.$$ 

• This produces margin of exactly 1 (why?)

• Again, the solution is expressed in terms of only a subset of examples.
  – These are the support vectors.
Lagrange multipliers

\[
\min_w \quad \frac{1}{2} \| w \|^2 = \frac{1}{2} \sum_{j=1}^{d} w_j^2,
\]

subject to \( y_i(w_0 + w^T x_i) - 1 \geq 0, \quad i = 1, \ldots, N. \)

- We will associate with each constraint the loss

\[
\max_{\alpha \geq 0} \alpha \left[ 1 - y_i(w_0 + w^T x_i) \right] =
\]
Lagrange multipliers

\[
\min_w \frac{1}{2} \parallel w \parallel^2 = \frac{1}{2} \sum_{j=1}^{d} w_j^2,
\]

subject to \( y_i(w_0 + w^T x_i) - 1 \geq 0, \quad i = 1, \ldots, N. \)

- We will associate with each constraint the loss

\[
\max_{\alpha \geq 0} \alpha \left[ 1 - y_i(w_0 + w^T x_i) \right] = \begin{cases} 
0, & \text{if } w_0 + w^T x_i - 1 \geq 0, \\
\infty & \text{otherwise (i.e. if constraint violated).}
\end{cases}
\]
Lagrange multipliers

\[
\min_w \frac{1}{2} \|w\|^2 = \frac{1}{2} \sum_{j=1}^{d} w_j^2,
\]

subject to \( y_i(w_0 + w^T x_i) - 1 \geq 0, \quad i = 1, \ldots, N. \)

- We will associate with each constraint the loss

\[
\max_{\alpha \geq 0} \alpha \left[ 1 - y_i(w_0 + w^T x_i) \right] = \begin{cases} 
0, & \text{if } w_0 + w^T x_i - 1 \geq 0, \\
\infty & \text{otherwise (i.e. if constraint violated)}. \end{cases}
\]

- We can reformulate our problem now:

\[
\min_w \frac{1}{2} \|w\|^2 + \sum_{i=1}^{N} \max_{\alpha_i \geq 0} \alpha_i \left[ 1 - y_i(w_0 + w^T x_i) \right]
\]
Max-margin optimization

- We want all the constraint terms to be zero:

\[
\min_w \left\{ \frac{1}{2} \| w \|^2 + \sum_{i=1}^{N} \max_{\alpha_i \geq 0} \alpha_i \left[ 1 - y_i (w_0 + w^T x_i) \right] \right\} \\
= \min_w \max_{\{\alpha_i \geq 0\}} \left\{ \frac{1}{2} \| w \|^2 + \sum_{i=1}^{N} \alpha_i \left[ 1 - y_i (w_0 + w^T x_i) \right] \right\}
\]
Max-margin optimization

• We want all the constraint terms to be zero:

\[
\begin{align*}
\min_w & \left\{ \frac{1}{2} \|w\|^2 + \sum_{i=1}^{N} \max_{\alpha_i \geq 0} \alpha_i \left[ 1 - y_i (w_0 + w^T x_i) \right] \right\} \\
= & \min_w \max_{\{\alpha_i \geq 0\}} \left\{ \frac{1}{2} \|w\|^2 + \sum_{i=1}^{N} \alpha_i \left[ 1 - y_i (w_0 + w^T x_i) \right] \right\} \\
= & \max_{\{\alpha_i \geq 0\}} \min_w \left\{ \frac{1}{2} \|w\|^2 + \sum_{i=1}^{N} \alpha_i \left[ 1 - y_i (w_0 + w^T x_i) \right] \right\}.
\end{align*}
\]
Max-margin optimization

- We want all the constraint terms to be zero:

\[
\min_w \left\{ \frac{1}{2} \|w\|^2 + \sum_{i=1}^{N} \max_{\alpha_i \geq 0} \alpha_i \left[ 1 - y_i (w_0 + w^T x_i) \right] \right\}
\]

\[
= \min_w \max_{\{\alpha_i \geq 0\}} \left\{ \frac{1}{2} \|w\|^2 + \sum_{i=1}^{N} \alpha_i \left[ 1 - y_i (w_0 + w^T x_i) \right] \right\}
\]

\[
= \max_{\{\alpha_i \geq 0\}} \min_w \left\{ \frac{1}{2} \|w\|^2 + \sum_{i=1}^{N} \alpha_i \left[ 1 - y_i (w_0 + w^T x_i) \right] \right\}. \quad J(w, w_0; \alpha)
\]

- We need to minimize \( J(w, w_0; \alpha) \) for any settings of \( \alpha = [\alpha_1, \ldots, \alpha_N]^T \).
Strategy for optimization

- We need to find
  \[
  \max_{\{\alpha_i \geq 0\}} \min_w J(w, w_0; \alpha)
  \]

- We will first fix \(\alpha\) and treat \(J(w, w_0; \alpha)\) as a function of \(w, w_0\).
  
  - Find functions \(w(\alpha), w_0(\alpha)\) that attain the minimum.

- Next, treat \(J(w(\alpha), w_0(\alpha); \alpha)\) as a function of \(\alpha\).
  
  - Find \(\alpha^*\) that attain the maximum.

- In the end, the solution is given by \(\alpha^*, w(\alpha^*)\) and \(w_0(\alpha^*)\).
Minimizing \( J(w, w_0; \alpha) \) with respect to \( w, w_0 \)

- For fixed \( \alpha \) we can minimize

\[
J(w, w_0; \alpha) = \frac{1}{2}||w||^2 + \sum_{i=1}^{N} \alpha_i \left[ 1 - y_i (w_0 + w^T x_i) \right]
\]

by setting derivatives w.r.t. \( w_0, w \) to zero:

\[
\frac{\partial}{\partial w} J(w, w_0; \alpha) = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0,
\]

\[
\frac{\partial}{\partial w_0} J(w, w_0; \alpha) = - \sum_{i=1}^{N} \alpha_i y_i = 0.
\]

- Note that the bias term \( w_0 \) dropped out but has produced a “global” constraint on \( \alpha \).
Solving for $\alpha$

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i y_i x_i, \quad \sum_{i=1}^{N} \alpha_i y_i = 0.$$ 

- Now can substitute this solution into

$$\max \left\{ \alpha_i \geq 0, \sum_i \alpha_i y_i = 0 \right\} \left\{ \frac{1}{2} \|w(\alpha)\|^2 + \sum_{i=1}^{N} \alpha_i \left[ 1 - y_i (w_0(\alpha) + w(\alpha)^T x_i) \right] \right\}$$

$$= \max \left\{ \alpha_i \geq 0, \sum_i \alpha_i y_i = 0 \right\} \left\{ \sum_{i=1}^{N} \alpha_i \left[ -\frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \right] \right\}.$$
Max-margin and quadratic programming

- We started by writing down the max-margin problem and arrived at the dual problem in $\alpha$:

$$\max \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \right\}$$

subject to $\sum_{i=1}^{N} \alpha_i y_i = 0$, $\alpha_i \geq 0$ for all $i = 1, \ldots, N$.

- Solving this quadratic program yields $\alpha^*$.

- We substitute $\alpha^*$ back to get $w$:

$$\hat{w} = w(\alpha^*) = \sum_{i=1}^{N} \alpha_i^* y_i x_i$$
Maximum margin decision boundary

\[ \hat{w} = w(\alpha^*) = \sum_{i=1}^{N} \alpha_i^* y_i x_i \]

• Suppose that, under the optimal solution, the margin of \( x_i \) is

\[ 1 - y_i \hat{w}^T x_i > 1. \]

• Then, necessarily, \( \alpha_i^* = 0 \Rightarrow \) not a support vector.

• We can then express the direction of the max-margin decision boundary

\[ \hat{w} = \sum_{\alpha_i^*>0} \alpha_i^* y_i x_i. \]

– We can compute \( w_0 \) by making sure the margin is balanced between the two classes (PS3).
Support vectors

\[ \hat{w} = \sum_{\alpha_i > 0} \alpha_i y_i x_i. \]

- Given a test example \( x \), it is classified by

\[
\hat{y} = \text{sign} \left( \hat{w}_0 + \hat{w}^T x \right) \\
= \text{sign} \left( \hat{w}_0 + \left( \sum_{\alpha_i > 0} \alpha_i y_i x_i \right)^T x \right) \\
= \text{sign} \left( \hat{w}_0 + \sum_{\alpha_i > 0} \alpha_i y_i x^T_i x \right)
\]

- The classifier is based on the expansion in terms of dot products of \( x \) with support vectors.
SVM classification

\[ \alpha > 0 \]

\[ \frac{1}{\|w\|} \]

\[ \alpha = 0 \]

\[ \alpha > 0 \]
SVM classification

\[ \alpha > 0 \]

\[ \frac{1}{\|w\|} \]

\[ \alpha = 0 \]

\[ \alpha > 0 \]

\[ \alpha > 0 \]

\[ \alpha > 0 \]
SVM: summary so far

- Assuming linearly separable case, we set up a quadratic program

\[
\max \left\{ \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j \right\}
\]

subject to \( \sum_{i=1}^{N} \alpha_i y_i = 0, \alpha_i \geq 0 \) for all \( i = 1, \ldots, N \).

- Solving it for \( \alpha \) we get the SVM classifier

\[
\hat{y} = \text{sign} \left( \hat{w}_0 + \sum_{\alpha_i > 0} \alpha_i y_i x_i^T x \right)
\]
Non-separable case

• What if the training data are non linearly separable? We can no longer require exact margin constraint.

• We re-write the constraints with *slack variables* $\xi_i$:

\[ y_i \left( w_0 + w^T x_i \right) - 1 + \xi_i \geq 0. \]

• The updated objective:

\[ \min_w \frac{1}{2} \| w \| + C \sum_{i=1}^{N} \xi_i. \]

• The parameter $C$ determines the penalty paid for violating the exact margin constraints.
Next time

10/13: Optimization.
10/16: no class.
10/18: Robot learning.
10/20: back to SVM: unseparable case, nonlinear classification.