An interesting dual - the SVM.

Problem: given N data points with 2 classes, \((x_i, y_i)\)

- wish to learn classifier for new points
- assume there is relatively little training data (otherwise we'd do something else)

Two cases:
- linearly separable
  - there is a hyperplane separating +ve or -ve examples
- Not linearly separable
  - there isn't
linearly separable:

- we want a linear classifier (so \( \text{sign}(a^T x + b) \) gives class)

- it must get training data right so \( y_i(a^T x + b) \geq 1 \), for each i

- if data is separable, we expect multiple \( a, b \) will work.

- Notice that if \( a, b \) are feasible, then so are \( \lambda a, \lambda b \), for \( \lambda > 0 \)

- Notice that the distance from \( x_i \) to the hyperplane \( a^T x_i + b \)

  \[
  \frac{|a^T x_i + b|}{\|a\|} \leq \sqrt{a^T a} \ ; \text{length of } a
  \]
This suggests choice of hyperplane s.t. all points are as far away as possible

\[ \min \frac{a^T a}{2} \]

\[ \text{st} \quad y_i (a^T x_i + b) \geq 1 \]

**Dual:**

\[ L_p (a, b, \lambda) = \frac{a^T a}{2} - \sum_i \lambda_i \left[ \left[ y_i (a^T x_i + b) - 1 \right]\right]^2 \]

\[ \nabla_a L = 0 = a - \sum_i \lambda_i [y_i x_i] \]

\[ \nabla_b L = 0 = -\sum_i \lambda_i y_i \]
Substitute these into Lagrangian
(notice coeff of b is \( \sum \lambda_i y_i = 0 \))
to get

\[
\max \sum_i \lambda_i - \frac{1}{2} \sum_{ij} \lambda_i \lambda_j [y_i y_j (x_i^T x_j)]
\]

\[
st: \sum_i \lambda_i y_i = 0 \quad (\text{inequality constraints})
\]

\(\lambda_i \geq 0\)
This problem is well studied.

- Consider constraints.
- Complementarity means that either
  \[ y_i (\mathbf{a}^T \mathbf{x}_i + b) = 1 \quad \text{OR} \quad \lambda_i = 0 \]

- Recall remarks on distance:
  all points where
  \[ y_i (\mathbf{a}^T \mathbf{x}_i + b) = 1 \]
  are the same distance from the hyperplane \( \mathbf{a} \mathbf{x} + b = 0 \)
  (but may be on different sides!)

- We do not expect many of these (geometry).
In $d$, $d+1$

This means almost every $\lambda_i$ is zero!

Obsolete algorithmic procedure:
- find the non-zero $\lambda_i$
- $a$ is then a weighted sum of those examples
- $b$ follows
Now assume problem isn't linearly separable.

This means
\[ y_i (a^T x_i + b) \geq 0 \]

\[ \rightarrow \text{feasible set} \quad \text{it's empty} \]

Introduce slacks
\[ y_i (a^T x_i + b) + \xi_i \geq 1 \]
\[ \xi_i \geq 0 \]

and penalize these slacks, to get
\[
\min \quad \frac{1}{2} a^T a + c \sum_i \xi_i \\
\text{st} \quad y_i (a^T x_i + b) + \xi_i \geq 1 \\
\xi_i \geq 0
\]
\[
\text{Dual:} \\
J = \frac{a^T a}{2} + c \sum \xi_i - \sum \lambda_i \left[ y_i (a^T x_i + b) - 1 + \xi_i \right] \\
- \sum \mu_i \xi_i
\]

\[
\nabla_a J = a - \sum \lambda_i y_i x_i = 0
\]

\[
\nabla_b J = 0 = -\sum \lambda_i y_i
\]

\[
\nabla \xi_i = c - \lambda_i - \mu_i = 0
\]

Subst back

\[
\text{Dual} = \sum \lambda_i - \frac{1}{2} \sum_{ij} y_i y_j x_i^T x_j \lambda_i \lambda_j
\]

st. \( \sum \lambda_i y_i = 0 \)

\( 0 \leq \lambda_i \leq C \) neat trick w/ \( \mu_i \)
This problem is well studied:

- Notice at soln, if $x_i$ is on the right side of boundary (i.e. correctly classified), $\mu_i > 0$ (complementarity!)

- It follows that for points far enough on the right side
  $\lambda_i = 0$, so $\mu_i = 0$
  $\uparrow$ complementarity

- So if most datapoints are correctly classified with high margin,
  most $\lambda_i = 0$

Obsolete algorithmic threat: find the non-zero $\lambda_i$. 
Notice we can interpret $\xi_i$ as a loss

$$\xi_i = \max(0, 1 - y_i(a^T x_i + b))$$

Chinge loss

$$h(a, b; x_i, y_i)$$

Then we have

$$\min_{a, b} \frac{a^T a}{2} + c \sum_i h(a, b; x_i, y_i)$$

(without constraints!)

Cleaner to write

$$\frac{1}{N} \sum_i h(a, b; x_i, y_i) + \lambda \frac{a^T a}{2}$$

\[\text{empirical hinge loss} \quad \text{regularizer}\]