Big Data Analytics

Special Topics for Computer Science
CSE 4095-001  CSE 5095-005

Feb 19

Fei Wang
Associate Professor
Department of Computer Science and Engineering
fei_wang@uconn.edu
Prediction II
Binary Classification

- \( f(x, w, b) = \text{sign}(w \cdot x + b) \)

- \( w \cdot x + b > 0 \) denotes +1
- \( w \cdot x + b < 0 \) denotes -1

How would you classify this data?
Binary Classification

\[ f(x, w, b) = \text{sign}(w \cdot x + b) \]

How would you classify this data?
Binary Classification

\[ f(x, w, b) = \text{sign}(w x + b) \]

How would you classify this data?
Binary Classification

\[ f(x, w, b) = \text{sign}(wx + b) \]

- denotes +1
- denotes -1

Any of these would be fine...

..but which is best?
Binary Classification

\[ f(x, w, b) = \text{sign}(w \cdot x + b) \]

• denotes +1
• denotes -1

How would you classify this data?

Misclassified to +1 class
The maximum margin linear classifier is the linear classifier with the maximum margin. This is the simplest kind of SVM (Called an LSVM).

1. Maximizing the margin is good according to intuition and PAC theory.
2. Implies that only support vectors are important; other training examples are ignorable.
3. Empirically it works very very well.

Support Vectors are those datapoints that the margin pushes up against.

Linear classifier with the, um, maximum margin.

This is the simplest kind of SVM (Called an LSVM).
Margin

- Distance from example to the separator is
  \[ r = y \frac{\mathbf{w}^T \mathbf{x} + b}{||\mathbf{w}||} \]

- Examples closest to the hyperplane are \textit{support vectors}.

- \textbf{Margin} \( \rho \) of the separator is the width of separation between support vectors of classes.

Derivation of finding \( r \):
Dotted line \( \mathbf{x}' - \mathbf{x} \) is perpendicular to decision boundary so parallel to \( \mathbf{w} \). Unit vector is \( \mathbf{w}/||\mathbf{w}|| \), so line is \( r\mathbf{w}/||\mathbf{w}|| \).
\( \mathbf{x}' = \mathbf{x} - yr\mathbf{w}/||\mathbf{w}|| \).
\( \mathbf{x}' \) satisfies \( \mathbf{w}^T \mathbf{x}' + b = 0 \).
So \( \mathbf{w}^T(\mathbf{x} - yr\mathbf{w}/||\mathbf{w}||) + b = 0 \)
Recall that \( ||\mathbf{w}|| = \sqrt{\mathbf{w}^T \mathbf{w}} \).
So \( \mathbf{w}^T\mathbf{x} - yr||\mathbf{w}|| + b = 0 \)
So, solving for \( r \) gives:
\[ r = y(\mathbf{w}^T \mathbf{x} + b)/||\mathbf{w}|| \]
Linear Support Vector Machine

- **Hyperplane**
  \[ w^T x + b = 0 \]

- **Extra scale constraint:**
  \[ \min_{i=1,...,n} |w^T x_i + b| = 1 \]

- This implies:
  \[ w^T(x_a - x_b) = 2 \]
  \[ \rho = \|x_a - x_b\|_2 = \frac{2}{\|w\|_2} \]
Linear Support Vector Machine

- Then we can formulate the *quadratic optimization problem*:

Find $\mathbf{w}$ and $b$ such that

$$\rho = \frac{2}{\|\mathbf{w}\|} \quad \text{is maximized; and for all } \{(\mathbf{x}_i, y_i)\}$$

$$\mathbf{w}^\top \mathbf{x}_i + b \geq 1 \text{ if } y_i = 1; \quad \mathbf{w}^\top \mathbf{x}_i + b \leq -1 \text{ if } y_i = -1$$

- A better formulation ($\min \|\mathbf{w}\| = \max 1/\|\mathbf{w}\|$):

Find $\mathbf{w}$ and $b$ such that

$$\Phi(\mathbf{w}) = \frac{1}{2} \mathbf{w}^\top \mathbf{w} \quad \text{is minimized;}$$

and for all $\{(\mathbf{x}_i, y_i)\}$: $y_i (\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$
Solving the Optimization Problem

Find \( w \) and \( b \) such that
\[
\Phi(w) = \frac{1}{2} w^T w \quad \text{is minimized;}
\]
and for all \( \{(x_i, y_i)\} \):
\[
y_i \left( w^T x_i + b \right) \geq 1
\]

- This is now optimizing a \textit{quadratic} function subject to \textit{linear} constraints
- Quadratic optimization problems are a well-known class of mathematical programming problem, and many (intricate) algorithms exist for solving them (with many special ones built for SVMs)
- The solution involves constructing a \textit{dual problem} where a \textit{Lagrange multiplier} \( \alpha_i \) is associated with every constraint in the primary problem:

Find \( \alpha_1, \ldots, \alpha_N \) such that
\[
Q(\alpha) = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x_i^T x_j
\]
is maximized and
\[
(1) \quad \sum \alpha_i y_i = 0 \\
(2) \quad \alpha_i \geq 0 \quad \text{for all} \quad \alpha_i
\]
The Solution

- The solution has the form:

\[ w = \sum \alpha_i y_i x_i \quad b = y_k - w^T x_k \] for any \( x_k \) such that \( \alpha_k \neq 0 \)

- Each non-zero \( \alpha_i \) indicates that corresponding \( x_i \) is a support vector.
- Then the classifying function will have the form:

\[ f(x) = \sum \alpha_i y_i x_i^T x + b \]

- Notice that it relies on an \textit{inner product} between the test point \( x \) and the support vectors \( x_i \)
  - We will return to this later.
- Also keep in mind that solving the optimization problem involved computing the inner products \( x_i^T x_j \) between all pairs of training points.
If the training data is not linearly separable, *slack variables* $\xi_i$ can be added to allow misclassification of difficult or noisy examples.

- Allow some errors
  - Let some points be moved to where they belong, at a cost
- Still, try to minimize training set errors, and to place hyperplane “far” from each class (large margin)
Soft-Margin SVM

- The old formulation:

\[
\text{Find } w \text{ and } b \text{ such that }
\Phi(w) = \frac{1}{2} w^T w \text{ is minimized and for all } \{(x_i, y_i)\}
\]
\[
y_i (w^T x_i + b) \geq 1
\]

- The new formulation incorporating slack variables:

\[
\text{Find } w \text{ and } b \text{ such that }
\Phi(w) = \frac{1}{2} w^T w + C \sum \xi_i \text{ is minimized and for all } \{(x_i, y_i)\}
\]
\[
y_i (w^T x_i + b) \geq 1 - \xi_i \text{ and } \xi_i \geq 0 \text{ for all } i
\]

- Parameter C can be viewed as a way to control overfitting
  - A regularization term
Soft-Margin SVM

- The dual problem for soft margin classification:

  Find $\alpha_1...\alpha_N$ such that
  
  $$Q(\alpha) = \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x_i^T x_j$$
  is maximized and 
  
  (1) $\sum \alpha_i y_i = 0$
  (2) $0 \leq \alpha_i \leq C$ for all $\alpha_i$

- Neither slack variables $\xi_i$ nor their Lagrange multipliers appear in the dual problem!
- Again, $x_i$ with non-zero $\alpha_i$ will be support vectors.
- Solution to the dual problem is:

  $$w = \sum \alpha_i y_i x_i$$
  $$b = y_k (1 - \xi_k) - w^T x_k$$
  where $k = \arg\max \alpha_k$.

  $w$ is not needed explicitly for classification!

  $$f(x) = \sum \alpha_i y_i x_i^T x + b$$
Prediction with SVM

- Given a new point $x$, we can score its projection onto the hyperplane normal:
  - I.e., compute score: $\mathbf{w}^T \mathbf{x} + b = \sum \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$
    - Decide class based on whether $< 0$ or $> 0$

- Can set confidence threshold $t$.
  
  Score $> t$: yes
  Score $< -t$: no
  Else: don’t know
Support Vector Machine: Linear Separable Case

- Assume that all data is at least distance 1 from the hyperplane, then the following two constraints follow for a training set \{ (x_i, y_i) \}

\[ w^T x_i + b \geq 1 \quad \text{if} \quad y_i = 1 \]
\[ w^T x_i + b \leq -1 \quad \text{if} \quad y_i = -1 \]

- For support vectors, the inequality becomes an equality
- Then, since each example’s distance from the hyperplane is

\[ r = y \frac{w^T x + b}{\|w\|} \]

- The margin is:

\[ \rho = \frac{2}{\|w\|} \]
Regression Analysis

In statistics, **regression analysis** is a statistical process for estimating the relationships among **variables**. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a **dependent variable** and one or more **independent variables**.

\[ y = ax + b \]

For the multi-dimensional case, the equation becomes:

\[ y = w^\top v + b \]

- **Features**
- **Outcome**
- **Slope**: 5/30
- **Offset**: 5

The graph shows a scatter plot with a linear regression line.
Least Square Regression

Prediction Loss

Single Dimensional Case

\[ \mathcal{L}(x) = (y - f(x))^2 \]

ax + b

\[
\begin{align*}
\mathcal{L}(\mathcal{X}) &= \sum_{i=1}^{n} (y_i - ax_i - b)^2 \\
&= \sum_{i=1}^{n} y_i^2 + a^2 \sum_{i=1}^{n} x_i^2 + nb^2 - 2a \sum_{i=1}^{n} y_i x_i - 2b \sum_{i=1}^{n} y_i - 2ab \sum_{i=1}^{n} x_i
\end{align*}
\]

\[ \nabla a \mathcal{L}(\mathcal{X}) = 2ra - 2s - 2pab = 0 \]

\[ \nabla b \mathcal{L}(\mathcal{X}) = 2nb - 2t - 2pab = 0 \]
LSR Example

Figure from http://abyss.uoregon.edu/~js/glossary/correlation.html
Binary Prediction

The following figure shows 30 day mortality in a sample of septic patients as a function of their baseline APACHE II Score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.

APACHE II ("Acute Physiology and Chronic Health Evaluation II") is a severity-of-disease classification system, one of several ICU scoring systems. It is applied within 24 hours of admission of a patient to an intensive care unit (ICU): an integer score from 0 to 71 is computed based on several measurements; higher scores correspond to more severe disease and a higher risk of death.

Sigmoid Function

\[ \pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \]

Probability Interpretation

death rate: \( \pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{1}{1 + \exp(-(\alpha + \beta x))} \)

survival rate: \( 1 - \pi(x) = \frac{1 - \exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} = \frac{1}{1 + \exp(\alpha + \beta x)} \)

odds of death: \( \frac{\pi(x)}{1 - \pi(x)} = \frac{\exp(\alpha + \beta x)}{1 - \exp(\alpha + \beta x)} \)

log odds of death: \( \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha + \beta x \)

logit function: for any number \(\pi\) between 0 and 1, \(\text{logit}(\pi) = \log(\pi/(1-\pi))\)
Logistic Loss

Let’s define an outcome $y_i$ for every $x_i$ such that
$y_i = 1$ if the $i$-th patient is dead
$y_i = -1$ if the $i$-th patient is survived

The likelihood of $y_i$ given the value of $x_i$

$$p(y_i | x_i) = \frac{1}{1 + \exp(-y_i(\alpha + \beta x_i))}$$

Under independence assumption, the likelihood of the training data set

$$p(Y | \mathcal{X}) = \prod_{i=1}^{n} p(y_i | x_i) = \prod_{i=1}^{n} \left( \frac{1}{1 + \exp(-y_i(\alpha + \beta x_i))} \right)$$

Maximum likelihood estimation is equivalent to minimize

$$\mathcal{L}(Y | \mathcal{X}) = \sum_{i=1}^{n} \log (1 + \exp(-y_i(\alpha + \beta x_i)))$$
MLE Estimation of Logistic Regression

\[
\frac{\partial L(y|x)}{\partial \alpha} = \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} \left( \log \left(1 + \exp \left(-y_i(\alpha + \beta x_i)\right) \right) \right)
\]

\[= \sum_{i=1}^{n} \frac{1}{1 + \exp \left(-y_i(\alpha + \beta x_i)\right)} \frac{\partial \left(\exp \left(-y_i(\alpha + \beta x_i)\right)\right)}{\partial \alpha} \]

\[= -\sum_{i=1}^{n} \frac{\exp \left(-y_i(\alpha + \beta x_i)\right)}{1 + \exp \left(-y_i(\alpha + \beta x_i)\right)} y_i
\]

\[= -\sum_{i=1}^{n} \frac{1}{1 + \exp \left(y_i(\alpha + \beta x_i)\right)} y_i \]