CSCI 181 / E-181 Spring 2014

2nd midterm review

David Wihl
davidwihl@gmail.com

April 26, 2014

1 Support Vector Machines

1.1 Background

Characteristics of SVMs:

- *stock* – SVMs are "off the shelf" and ready to use. No special modification is necessary.
- *linearly separable* – assumes that linear separation is possible. Used natively as a binary classifier.
- *convex optimization*. SVM originated as a backlash against neural nets due to nets' non-convexity. In Neural Nets, results were often non-reproducible as different researchers found different results due to different initializations.
- *global optimum* – SVMs will find the global optimum.

SVMs are based on three "big ideas":

- *margin* Maximizes distance between the closest points
- *duality* Take a hard problem and transform it into an easier problem to solve.
- *kernel trick* Map input vectors to higher dimensional, more expressive features.

1.2 Definitions

Data: $\{x_n, t_n\}_{n=1}^{N}, t_n \in \{-1, +1\}$. t_n is the target or the expected result of the classification.

J Basis functions: $\phi_j(x) \to \Re$, therefore

Vector function: $\Phi X \to \Re^J$ produces a column vector.

Objective function: $f(\mathbf{x}, \mathbf{w}, b) = \mathbf{w}^\intercal \phi(\mathbf{x}) + b$ where b is the bias.

The sign of $f(\cdot)$ will determine classification $(-1, +1)$

So the actual classifier will be:

$$y(\mathbf{x}, \mathbf{w}, b) = \begin{cases} +1, & \text{if } \mathbf{w}^\intercal \phi(\mathbf{x}) + b > 0 \\ -1, & \text{otherwise} \end{cases}$$

Unlike Logistic Regression (which uses $\{0, 1\}$), it is preferable to use $\{-1, +1\}$ as the classification result. If $t_n * y$ is positive, then the produced classification is correct (positive $\times$ positive is positive, negative $\times$ negative is also positive).

Decision Boundary is the hyperplane where $\mathbf{w}^\intercal \phi(\mathbf{x}) + b = 0$. We want to find the Decision Boundary that creates the most separation between the two different classes by maximizing the distance between the two closest points. The distance between the Decision Boundary and the closest point is called the *margin*. The points closest to the Decision Boundary are called the *support vectors*.

1.3 Max Marginalization

The margin is determined by the orthogonal distance from the closest point to the Decision Boundary:

$$\frac{|\mathbf{w}^\intercal \mathbf{x} + b|}{||\mathbf{w}||} \tag{1}$$

Maximizing the margin can be written as:

$$\underset{w,b}{\text{argmax}} \left\{ \min_n (t_n \cdot (\mathbf{w}^\intercal \mathbf{x} + b)) \cdot \frac{1}{||w||} \right\} \tag{2}$$

Maximizing the margin helps ensure that points which are close to margin will not be pushed over the boundary by noise.

w is orthogonal to vectors in the Decision Boundary. Here's how: pick two points on the

Decision Boundary $\phi(x_1)$ and $\phi(x_2)$. So

$$\mathbf{w}^\intercal (\phi(x_2) - \phi(x_1)) = 0 \text{ for orthogonal dot product}$$
$$\mathbf{w}^\intercal \phi(x_2) - \mathbf{w}^\intercal \phi(x_1) = 0$$

Note: $\phi(x_n) = (-b)$
$=(-b) - (-b)$
$=0$

1.4 Duality

1.5 Kernel Tricks

Mercer function, infinite dimensions (justification for duality)

Slack variables to break the linear separability.

1.6 Sources

1. Lecture 14, March 24, 2014
2. Lecture 15
3. Bishop 6.0-6.2
4. Bishop 7.0-7.1
5. Course notes - maxmargin
6. Section 7 review
7. Section 8 review
8. Machine Learning in Action, Chapter 6

2 Markov Decision Processes

Lecture 16

Course notes - MDP

Section 9

2.1 Partially Observable MDP

Course notes - POMDP

Section 10

2.2 Hidden Markov Models

Bishop 13.0-13.2

2.3 Mixture Models

Bishop 9.0-9.2

3 Reinforcement Learning

Course notes - RL

Section 9

3.1 Value and Policy Iteration

Lecture 17

Course notes - policyiter

4 Expectation Maximization

Bishop 9.3

Section 11