CS 446: Machine Learning
Discussion Session

Daniel Khashabi

November 6, 2015

1 Support Vector Machines:

Consider a dataset with 3 points in 1-D:

$$\{(+,0),(-,-1),(-,+1)\}$$

1. Are the classes $\pm$ linearly separable?
2. Consider mapping each point to 3D using new feature vectors $\Phi(x) = \left[1, x\sqrt{2}, x^2\right]^\top$. Are the classes now linearly separable? If so, find a separating hyperplane.
3. Consider the formulation of the soft-margin primal SVM, for a given training data:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\ y_i \in \{-1,1\}\}_{i=1}^n$$

$$\arg\min_{\mathbf{w},\xi,b} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i \right\}$$
$$y_i(\mathbf{w}\cdot\mathbf{x_i} - b) \geq 1 - \xi_i, \quad \forall i = 1,..,n$$
$$\xi_i \geq 0, \quad \forall i = 1,..,n$$

Also remember the hard-margin primal SVM $\sigma_i = 0, \forall i$. And remember that we can derive the dual formulation and replace each $\mathbf{x}.\mathbf{x}'$ with a kernel function $k(\mathbf{x},\mathbf{x}')$.

Mach each of the followings with a decision boundary in Figure 1:

(a) A soft-margin linear SVM with $C = 0.1$.
(b) A soft-margin linear SVM with $C = 10$.
(c) A hard-margin kernel SVM with kernel $k(\mathbf{u},\mathbf{v}) = \mathbf{u}.\mathbf{v} + (\mathbf{u}.\mathbf{v})^2$.
(d) A hard-margin kernel SVM with kernel $k(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{1}{4}\|\mathbf{u}-\mathbf{v}\|^2\right)$.
(e) A hard-margin kernel SVM with kernel $k(\mathbf{u},\mathbf{v}) = \exp\left(-4\|\mathbf{u}-\mathbf{v}\|^2\right)$.

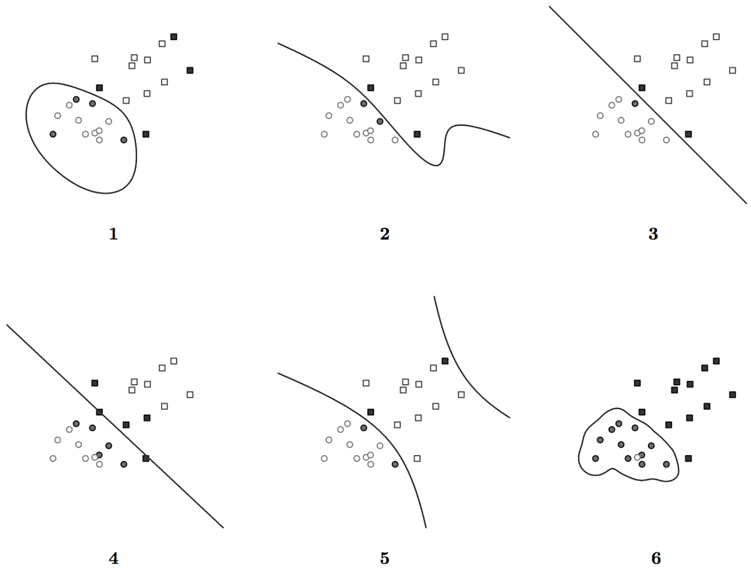

Figure 1: Decision Boundaries

4. Define a class variable $y_i \in \{-1, +1\}$ which denotes the class of x_i and let $\mathbf{w} = (w_1, w_2, w_3)^\top$. The max-margin SVM classifier solves the following problem

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
$$y_i(\mathbf{w} \cdot \phi(\mathbf{x_i}) - b) \geq 1, \quad \forall i = 1, .., n$$

Using the method of Lagrange multipliers show that the solution is $\hat{\mathbf{w}} = (0, 0, -2)^\top$, $b = 1$ and the margin is $1/\|\hat{\mathbf{w}}\|$.

5. What happens if we change the constraints to

$$y_i(\mathbf{w} \cdot \phi(\mathbf{x_i}) - b) \geq \beta, \beta \geq 1$$

Solution:

1. No.
2. The points are mapped to $(1, 0, 0)$, $(1, -\sqrt{2}, 1)$, $(1, \sqrt{2}, 1)$, respectively. The points are now separable in 3-dimensional space. A separating hyperplane is given by the weight vector $(0, 0, 1)$.
3. ?

4. First notice that all of the three points are support vectors. Therefore:

$$\arg\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2$$
$$y_i(\mathbf{w}\cdot\phi(\mathbf{x_i}) - b) = 1, \quad \forall i = 1,2,3$$
$$L(\mathbf{w},\alpha) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1,2,3} \alpha_i\left(y_i(\mathbf{w}\cdot\phi(\mathbf{x_i}) - b) - 1\right)$$
$$\frac{\partial L(\mathbf{w},\alpha)}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1,2,3} \alpha_i y_i \phi(\mathbf{x_i}) = 0$$
$$\frac{\partial L(\mathbf{w},\alpha)}{\partial b} = \sum_{i=1,2,3} \alpha_i y_i = 0$$

Therefore:

$$w_1 + \alpha_1 - \alpha_2 - \alpha_3 = 0$$
$$w_2 + \sqrt{2}\alpha_2 - \sqrt{2}\alpha_3 = 0$$
$$w_3 + \alpha_2 - \alpha_3 = 0$$
$$\alpha_1 - \alpha_2 - \alpha_3 = 0$$

which would give us the desired result.

5. ?

2 Probabilistic Estimation:

In this problem we will find the maximum likelihood estimator (MLE) and maximum a posteriori (MAP) estimator for the mean of a univariate normal distribution. Specifically, we assume we have N samples, $x_1, ..., x_N$ independently drawn from a normal distribution, with *unknown* mean μ and *known* variance σ^2.

1. Derive the MLE estimator for the mean μ.
2. Suppose we have a Gaussian prior on μ, with mean ν and variance β^2. Derive the MAP estimation for μ.
3. Comment on what happens to the MLE and MAP estimators as the number of samples N goes to infinity.

Solution:

1.

$$\begin{aligned} P(x_1, ..., x_N|\mu, \sigma^2) &= \prod_i P(x_i|\mu, \sigma^2) \\ &= \prod_i \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{\sum_i (x_i-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

Now we form the log-likelihood functions:

$$L = \log P(x_1, ..., x_N|\mu, \sigma^2) = n \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2}$$

$$\frac{\partial L}{\partial \mu} = 0 \Rightarrow \mu_{ML} = \frac{\sum_i x_i}{n}$$

2. Using the Bayes rule:

$$P(\mu|x_1, ..., x_N, \sigma^2) = \frac{P(x_1, ..., x_N|\mu, \sigma^2)P(\mu)}{P(x_1, ..., x_N)}$$

The target is to find:

$$\mu_{MAP} = \arg\max_{\mu} P(\mu|x_1, ..., x_N, \sigma^2) = \arg\max_{\mu} P(x_1, ..., x_N|\mu, \sigma^2)P(\mu)$$

Now we simplify the RHS:

$$P(x_1, ..., x_N|\mu, \sigma^2)P(\mu) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{\sum_i (x_i - \mu)^2}{2\sigma^2}\right\} \times \frac{1}{\beta\sqrt{2\pi}} \exp\left\{-\frac{(\mu - \nu)^2}{2\beta^2}\right\}$$

And taking the log from both sides:

$$\Gamma = \ln\left[P(x_1, ..., x_N|\mu, \sigma^2)P(\mu)\right] = C - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2} - \frac{(\mu - \nu)^2}{2\beta^2}$$

Taking derivative with respect to μ:

$$\frac{\partial \Gamma}{\partial \mu} = 0 \Rightarrow \mu_{MAP} = \frac{\sigma^2\nu + \beta^2 \sum_i x_i}{\sigma^2 + n\beta^2}$$