Statistics Review

junyan.xu*

April 8, 2019

Contents

1 Support Vector Machine and Regression **2**
1.1 Separable 2
1.2 Non Separable 2
1.3 Solver 2
1.4 Dual Problem 2
1.5 Regression 3
1.6 Summary 3

2 Kalman Filter and Hidden Markov Model **3**
2.1 Basic Definition 3
2.2 EM Algorithm 3

*junyanxu5513@gmail.com

1 Support Vector Machine and Regression

1.1 Separable

A distance of a point a to a plane $w \cdot x + b = 0$ is $\frac{a \cdot w + b}{\|w\|}$. Define margin ζ to be the value that all (x_i, y_i) are having distance larger that this.

$$\frac{x_i \cdot w + b}{\|w\|} \geq \zeta \tag{1}$$

Then if we fix ζ to be 1, mathematically the margin is $\frac{1}{\|w\|}$. The we are actually minimize $\frac{1}{2}\|w\|^2$

1.2 Non Separable

If the data is not separable, then we add penalty to the optimization function (constant muliply soft margin).

$$\begin{aligned} loss &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \\ s.t. \forall i, &y_i(w.x_i + b) >= 1 - \xi_i \end{aligned} \tag{2}$$

The x_i represent the hinge loss, an "option" style loss function. Which could be rewrite as $max\{0, 1 - y(w \cdot x + b)\}$

1.3 Solver

Currently according to Stanford course this one should use SGD to solve.

1.4 Dual Problem

According to representer theorem the w can be a linear combination of x_i and is $w = \sum_{i=1}^{n} \alpha_i x_i$. Then we have $f(x_i) = y_i \sum_{j=1}^{n} \alpha_j x_j^T x_i + b$ The primal problem of this is minimize:

$$\frac{1}{2}\sum_{jk} \alpha_j \alpha_k y_j y_k (x_j^\top) + \sum_{i=1}^{n} C \cdot max(0, 1 - y_i \sum_{j=1}^{n} \alpha_j x_j^T x_i + b) \tag{3}$$

The dual problem is maximize

$$\begin{aligned} &-\frac{1}{2}\sum_{jk} \alpha_j \alpha_k y_j y_k (x_j^\top) + \sum_{i=1}^{n} \alpha_i \\ &0 \leq \alpha_i \leq C \\ &\sum_i \alpha_i y_i = 0 \end{aligned} \tag{4}$$

1.5 Regression

For regression part. The we assume the regression lies in range ϵ. Also the penalty of upper ϵ is ξ_u and down is ξ_d So the primal problem is

$$loss = \frac{1}{2}\|w\|^2 + C\sum_{i=1}(max\{f(x_i) - y_i - \epsilon, 0\} + max\{y_i - f(x_i) - \epsilon, 0\}) \tag{5}$$

1.6 Summary

1. GOOD: SVM is good for handling high dimensional data
2. GOOD: SVM has customize kernal
3. BAD: No probability estimate

2 Kalman Filter and Hidden Markov Model

2.1 Basic Definition

Kalman filter has definition x_t and the transition matrix A_t. The x_t follows the formula

$$\begin{aligned} x_{t+1} &= A_t \cdot x_t + Normal(0, Q_t) + b_t \\ z_t &= C_t \cdot x_t + Normal(0, R_t) + d_t \end{aligned} \tag{6}$$

All the state transitions and observations are linear with Gaussian distributed noise, then the estimation can be represented by a mean plus a Gaussian distribution.

The kalman gain is a critical term in Kalman Filter. In the formula above, kalman gain can be defined as the coeffcient of innovation error that we need to update our prior estimation of x_t based on the current observation z_t. The idea is quite straight forward

$$\hat{x}_t|t = \hat{x}_t|t-1 + K_t \cdot (z_t - C_t \cdot \hat{x}_t|t-1)$$

The solving of kalman gain can be defined as minimize covariance of time t posterior estimation of x. Which is actually miniize the trace of error matrix.

$$K_t = \arg\min_K trace(cov(x_t - x_t|t-1 - K(C \cdot x_t + v_t - C_t \cdot x_t|t-1))) \tag{7}$$

2.2 EM Algorithm

Two step, the E step of this algorithm first assumed we have θ, then we can get distribution of hidden x. Then we calculate the expectation $E_x(l)$ of the likelyhood under hidden variable$x$. Next step is the M step, caclculate the max E to get theta