$$L(w,b,\varepsilon,\Lambda)=\frac{1}{2}w^Tw+\frac{c}{2}\sum_{i=1}^{l}\varepsilon_i^2-\sum_{i=1}^{l}\alpha_i\{y_i(w^Tx_i+b)-1+\varepsilon_i\}-\sum_{i=1}^{l}r_i\varepsilon_i$$

$$\frac{\partial L}{\partial w}=0\rightarrow \boxed{w-\sum_{i=1}^{l}\alpha_i y_i x_i=0}^{(*)},\quad \frac{\partial L}{\partial b}=0\rightarrow \boxed{\sum_{i=1}^{l}\alpha_i y_i=0}^{(*)}$$

$$\frac{\partial L}{\partial \varepsilon_i}=0\rightarrow c\varepsilon_i-\alpha_i-r_i=0\rightarrow \boxed{c\varepsilon_i=r_i+\alpha_i}^{(\heartsuit)}$$

$$\boxed{r_i\varepsilon_i=0}^{(+)},\quad \boxed{\alpha_i\{y_i(w^Tx_i+b)-1+\varepsilon_i\}=0}$$

$$\Rightarrow L(w,b,\varepsilon,\Lambda)=\frac{1}{2}w^Tw+\frac{c}{2}\sum_{i=1}^{l}\varepsilon_i^2-\underbrace{\sum_{i=1}^{l}\alpha_i y_i(w^Tx_i)}_{-w^Tw}-\underbrace{\sum_{i=1}^{l}\alpha_i y_i b}_{0}+\sum_{i=1}^{l}\alpha_i$$

(w طبق رابطه بالای (*)، 0 طبق (*))

$$-\sum_{i=1}^{l}\alpha_i\varepsilon_i-\sum_{i=1}^{l}r_i\varepsilon_i=-\sum_{i=1}^{l}\underbrace{(\alpha_i+r_i)}_{c\varepsilon_i\ \text{طبق}\ (\heartsuit)}\varepsilon_i=-\sum_{i=1}^{l}c\varepsilon_i^2$$

$$\rightarrow =\frac{1}{2}w^Tw+\frac{c}{2}\sum_{i=1}^{l}\varepsilon_i^2-w^Tw+\sum_{i=1}^{l}\alpha_i-\sum_{i=1}^{l}c\varepsilon_i^2$$

$$=\boxed{-\frac{1}{2}w^Tw-\frac{c}{2}\sum_{i=1}^{l}\varepsilon_i^2+\underbrace{\sum_{i=1}^{l}\alpha_i}_{c\sum_{i=1}^{l}\varepsilon_i}}^{(\bullet)}$$

$$(\heartsuit)\rightarrow c\varepsilon_i=r_i+\alpha_i\rightarrow r_i=c\varepsilon_i-\alpha_i$$
$$(+)\rightarrow r_i\varepsilon_i=0\xrightarrow{\div\varepsilon_i} r_i=0$$
$$\Big\}\ c\varepsilon_i-\alpha_i=0\Rightarrow \boxed{\alpha_i=c\varepsilon_i}$$

$$\xrightarrow{(\bullet)}\boxed{L(w,b,\varepsilon,\Lambda)=-\frac{1}{2}w^Tw-\frac{c}{2}\sum_{i=1}^{l}\varepsilon_i^2+c\sum_{i=1}^{l}\varepsilon_i}$$

Subject: Date:

6- الف) در روش های discriminative ، مدل از احتمال شرطی $P(w_i|x)$ برای پیش بینی داده های unseen استفاده می کند. این روش ها برای بدست آوردن احتمال، به طور مستقیم یک توزیع فرضی برای $P(w_i|x)$ درنظر می‌گیرند و سپس پارامترهای این توزیع را با کمک داده های train تخمین می‌زنند. مدل های discriminative در واقع مرز تصمیم بین کلاس های مختلف را مدل می کنند. برخی مدل های discriminative عبارتند از: Logistic Regression ، SVMs ، nearest neighbor ، decision tree & Random Forest و ... . این مدل ها قابلیت تولید داده‌ی جدید ندارند و به طور کلی هدف این مدل ها، جداسازی یک کلاس داده ها از دیگری است.

روش های Generative ، برای بدست آوردن احتمال شرطی $P(Y|X)$ ، احتمال prior $P(Y)$ و likelihood $P(X|Y)$ را به کمک داده‌ی train تخمین می زنند و سپس طبق رابطه‌ی بیز،

احتمال joint (یعنی $P(x,y)$) را بدست می‌آورند و این احتمال joint می‌تواند به عنوان جایگزینی برای احتمال شرطی $P(Y|X)$ استفاده شود تا به کمک آن پیش بینی انجام دهیم.

$$P(x,y) = P(y) \cdot P(x|y)$$

برخی مدل های generative عبارتند از: Bayesian networks ، Hidden Markov model ، GANs و... این مدل ها قابلیت تولید داده‌ی جدید دارند و به طور کلی هدف آن ها مدلسازی خود داده هاست (برخلاف مدل های discriminative که هدف آن ها مدلسازی مرز تصمیم بین کلاس های مختلف بود)

در کل، از آنجایی که (همانطوری که توضیح دادم) بدست آوردن $P(x,y)$ بسیار زمانبر است، روش های discriminative در مقابل روش های generative ، efficient تر هستند اما نسبت به روش های generative ، دقت کمتری هم دارند.

ب) مهمترین مزیت روش one vs. all این است که مسأله‌ی طبقه بندی داده های c کلاس را به c تا مسئله‌ی binary classification تبدیل می‌کند (تبدیل مسئله‌ی multi_class classification به تعدادی مسئله‌ی ساده‌تر binary classification). اما عیب این روش این است که مجبوریم برای هر کلاس از داده ها، یک مدل بسازیم. ساختن تعداد زیادی مدل مخصوصاً برای زمانی که دیتاست بزرگی داریم و زمانی که تعداد کلاس های داده‌ی ما زیاد است، با استفاده از مدل های کندی مثل شبکه های عصبی، بسیار چالش برانگیز خواهد بود. یکی دیگر از معایب این روش این است که ممکن است توزیع داده های کلاس در هریک از این مسائل binary classification ، نامتقارن شود و در چنین حالتی طبقه بندها نیز عملکرد خوبی برای کلاسی که داده های آن از بقیه کمتر و آشفته‌تر است نداشته باشند.

Subject: Date:

۲) بهترین مزیت dual نسبت به primal زمانی است که می‌خواهیم از kernel برای classify کردن داده هایی استفاده کنیم که در فضای اولیه به صورت خطی قابل جداسازی نیستند. dual form به راحتی با استفاده از kernel قابل اعمال است. در چنین حالتی (وقتی از kernel استفاده می‌کنیم)، مسئله‌ی optimization توسط فرم dual به صورت زیر می‌باشد:

$$\max \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j)$$

$$s.t. \quad \forall i \ \alpha_i \geq 0 \ \wedge \ \sum_{i=1}^{n} y_i \alpha_i = 0$$

می‌بینیم که این فرم تقریباً با فرم اولیه‌ی dual یکسان است. تنها تفاوت این است که به‌جای dot product عادی، باید تابع kernel را محاسبه کنیم. دقت کنید که چنین چیزی در فرم primal ممکن نیست. چرا که باید صریحاً mapping هر نقطه را محاسبه کنیم:

$$\min_{w, w_0} \frac{1}{2} w^T w$$

$$s.t. \quad \forall i: \ y_i (w^T \phi(x_i) + w_0) \geq 1$$

همچنین بازه‌ی dual، classify کردن یک نقطه‌ی جدید ساده‌تر خواهد بود.

در کل با استفاده از روش dual، داده هایی که به صورت خطی قابل جداسازی نیستند را به راحتی می‌توان classify کرد (با کمی پیچیدگی محاسباتی بیشتر). (اما برای linear SVM، حل کردن فرم primal ساده‌تر است)

Subject: Date:

7_

الف)

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2}\right) = \varphi^t(x_i)\,\varphi(x_j)$$

$$d(\varphi(x_i), \varphi(x_j)) = \|\varphi(x_i) - \varphi(x_j)\| = \sqrt{(\varphi(x_i) - \varphi(x_j))^t(\varphi(x_i) - \varphi(x_j))}$$

$$= \sqrt{\underbrace{\varphi^t(x_i)\varphi(x_i)}_{K(x_i,x_i)} - \underbrace{\varphi^t(x_i)\varphi(x_j)}_{K(x_i,x_j)} - \underbrace{\varphi^t(x_j)\varphi(x_i)}_{=\varphi(x_i)^t\varphi(x_j)=} + \underbrace{\varphi^t(x_j)\varphi(x_j)}_{K(x_j,x_j)}}$$

$$= \sqrt{K(x_i,x_i) - 2K(x_i,x_j) + K(x_j,x_j)} =$$

($e^0 = 1$ for both $K(x_i,x_i)$ and $K(x_j,x_j)$)

$$= \sqrt{2 - 2K(x_i,x_j)} = \sqrt{2 - 2\exp\left(-\frac{\|x_i - x_j\|^2}{2}\right)} \leqslant \sqrt{2}$$

$$\Rightarrow \|\varphi(x_i) - \varphi(x_j)\|^2 \leqslant (\sqrt{2})^2 = 2$$

ب)

1) $K(x,y) = f(x)\,K_1(x,y)\,f(y)$

طبق خواص kernel ها می دانیم: $K(x,y) = f(x).f(y)$ برای هر تابع f روی x، زیرا فقط

یک ویژگی توسط f، define می شود. پس می توانیم بنویسیم: $K_2(x,y) = f(x)f(y)$ (*)

همچنین طبق خواص kernel ها می دانیم: $K(x,y) = K_1(x,y)K_2(x,y)$

پس می توانیم بنویسیم: $K(x,y) = K_1(x,y)K_2(x,y) \overset{(*)}{=} K_1(x,y)f(x)f(y)$