به نام خدا

دانشکده‌ی مهندسی کامپیوتر

سررسید تئوری: ۲۳ آذر ماه پنجشنبه ۲۳:۵۹
سررسید عملی: ۲۵ آذر ماه شنبه ۲۳:۵۹ پاییز ۱۴۰۲

یادگیری ماشین

مدرس: مهدی جعفری سیاوشانی تمرین ۴: SVM & Kernel Methods

- سررسید بخش تئوری این تمرین پنجشنبه ۲۳ آذر ماه ساعت ۵۹ : ۲۳ است.
- سررسید بخش عملی این تمرین شنبه ۲۵ آذر ماه ساعت ۵۹ : ۲۳ است.
- در صورت کشف تقلب، بار اول برای افراد درگیر تقلب، نمره‌ی همان سوال(های) خاص صفر در نظر گرفته می‌شوند. در صورت تکرار، نمره کل تمرین صفر در نظر گرفته می‌شود و در صورت تکرار، درس برای افراد حذف خواهد شد.
- تمامی پاسخ‌های خود را در یک فایل با فرمت (.pdf) HW4-[SID]-[Fullname].zip روی کوئرا قرار دهید.

پرسش‌ها

۱ قسمت تئوری

۱.۱ پرسش اول (۲۰ نمره)

۱. (۷ نمره) می‌دانیم ویژگی‌های زیر برای هر هسته معتبر دلخواه مانند $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ برقرار است:

(آ) تقارنی $\leftarrow \forall x, y \in \mathbb{R}^d : f(x, y) = f(y, x)$

(ب) افزایشی (بر حسب اولین آرگومان) $\leftarrow \forall x, y, z \in \mathbb{R}^d : f(x + z, y) = f(x, y) + f(z, y)$

(ج) همگنی (بر حسب اولین آرگومان) $\leftarrow \forall x, y \in \mathbb{R}^d, \alpha \in \mathbb{R}_{++} : f(\alpha x, y) = \alpha f(x, y)$

با این فرض که هسته $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ هر سه ویژگی بالا را دارد. نشان دهید هسته $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ که به شکل زیر تعریف می‌شود معتبر است.

$$h(x, y) = \frac{1}{4}(g(x + y, x + y) - g(x - y, x - y))$$

۲. (۱۰ نمره) با فرض معتبر بودن هسته‌های k_1 و k_2 ، اعتبار هسته‌های زیر را بررسی کنید.

(آ) $k_3(x_1, x_2) = k_1(x_1, x_2) + k_2(x_1, x_2)$

(ب) $k_4(x_1, x_2) = k_1(x_1, x_2)k_2(x_1, x_2)$

(ج) $k_5(x_1, x_2) = e^{k_1(x_1, x_2)}$

(د) $k_6(x_1, x_2) = (1 - x_1^T x_2)^{-1}$

۳. (۳ نمره) $\hat{X}$ را مجموعه تمام زیرمجموعه‌های متناهی X در نظر بگیرید. ثابت کنید اگر K یک کرنل معتبر روی $X \times X$ باشد آنگاه $\hat{k}(A, B) = \sum_{x \in A, x' \in B} k(x, x')$ یک کرنل معتبر روی $X' \times X'$ است.

۲.۱ پرسش دوم (نمره ۱۵)

۱. (۶ نمره) در اسلایدهای درس ثابت شد که فرم دوگان مسئله‌ی SVM به صورت بیشینه کردن

$$\tilde{L}(a) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1, m=1}^{N} a_n a_m y^{(n)} y^{(m)} k(x^{(n)}, x^{(m)})$$

با شروط

$$\forall a_i \geqslant 0 \quad \& \quad \sum_{n=1}^{N} a_n y^{(n)} = 0$$

است. نشان دهید اگر a جواب بهینه باشد، میزان حاشیه برابر است با $\rho = \frac{1}{\sqrt{\sum_{n=1}^{N} a_n}}$.

۲. (۳ نمره) در hard margin SVM فرض کنید که یک ویژگی بی ربط اضافه کنیم که تاثیری در افزایش margin ندارد. توضیح دهید که آیا SVM نسبت به چنین ویژگی robust است یا خیر و اینکه چگونه در این سناریو عمل می‌کند؟

۳. (۶ نمره) فرض کنید داده‌های ما در یک فضای d بعدی هستند $(d > ۲)$. نشان دهید که مجموعه دادگان آموزش شامل فقط دو داده با برچسب های متفاوت برای تعیین فاصله‌ی ابرصفحه جداکننده از مرکز مختصات کافی است.

۳.۱ پرسش سوم (۲۰ نمره)

در کلاس درس درباره استفاده از SVM ها برای دسته‌بندی بحث شد. حال در این سوال قصد داریم این مفهوم را به مسئله رگرسیون انتقال دهیم. برای این منظور، همان روالی که انجام شد را مرحله به مرحله انجام می‌دهیم. فرض کنید داده‌هایتان $(x^{(۱)}, y^{(۱)}) \cdots (x^{(n)}, y^{(n)})$ باشند که $x^{(i)} \in \mathbb{R}^d$ ، $y^{(i)} \in \mathbb{R}$. یک تابع ضرر متداول برای این منظور به صورت

$$L_\epsilon(x, y, f) = |y - f(x)|_\epsilon = max(۰, |y - f(x)| - \epsilon)$$

است که با بکارگیری آن، تابع هزینه کلی به صورت زیر درمی‌آید.

$$\frac{۱}{۲}\|w\|^۲ + C \sum_{i=۱}^{n} L_\epsilon(x^{(i)}, y^{(i)}, f)$$

۱. (۲ نمره) با تعریف متغیر slack (ξ_i و ξ_i^*) و اعمال شرط مناسب روی آن، نشان دهید صورت اصلی (primal) این مسئله (که یک مسئله quadratic است) به شکل

$$\min_{w \in \mathbb{R}^m, \xi \in \mathbb{R}^n, \xi^* \in \mathbb{R}^n} \frac{۱}{۲}\|w\|^۲ + C \sum_{i=۱}^{n} (\xi_i^* + \xi_i)$$

می‌باشد. (راهنمایی: همانطور که خواندید، در soft margin SVM متغیر ξ_i میزان تخطی حاشیه را نشان می‌داد. در این مسئله ξ_i را میزان تخطی مربوط به بیشتر بودن پیشبینی از y_i و ξ_i^* را میزان تخطی کمتر بودن پیشبینی از y_i بگیرید.)

۲. (۸ نمره) همانند روال اسلاید ابتدا تابع لاگرانژین برای صورت اصلی بنویسید سپس با عوض کردن ترتیب max و min و استفاده از شرط‌های K.K.T به صورت دوگان برسید.

$$\max_{\alpha \in \mathbb{R}^n, \alpha^* \in \mathbb{R}^n} -\frac{۱}{۲} \sum_{i,j=۱}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j \rangle - \epsilon \sum_{i=۱}^{n} (\alpha_i + \alpha_i^*) + \sum_{i=۱}^{n} y_i(\alpha_i - \alpha_i^*)$$

با شرط

$$\alpha_i, \alpha_i^* \in [۰, C].$$

۳. (۱ نمره) توضیح دهید که آیا صورت دوگان مسئله با quadratic optimization solver قابل حل است؟

۴. (۲ نمره) در این مسئله support vector ها به چه صورت مشخص می‌شوند؟

۵. (۳ نمره) رابطه‌ای برای پیشبینی داده‌ی جدید بنویسید و توضیح دهید که آیا می‌شود از تکنیک کرنل استفاده کرد؟

۶. (۴ نمره) تغییر ϵ موجب چه می‌شود؟ تغییر C چطور؟

۴.۱ پرسش چهارم (نمره ۱۵)

۱. (۲ نمره) اگر در یک مسئله خطی تفکیک پذیر (با روش حل SVM) یکی از دادگان آموزش حذف شود، مرز تصمیم به سمت نقطه حذف شده جابجا می‌شود یا در خلاف جهت آن؟ یا اصلا تغییری نمی‌کند؟ توضیح دهید. حال اگر فرض کنیم مرز تصمیم [1] برای Logistic Regression است؛ مرز تصمیم جابجا می‌شود یا ثابت باقی می‌ماند؟ توضیح دهید. (نیازی به مشخص کردن جهت تغییر نیست)

۲. (۷ نمره) با توجه به درس اگر اجازه تعدادی دسته‌بندی اشتباه در دادگان آموزش بدهیم، بهینه‌سازی اصلی (primal) ، (soft margin SVM) به صورت زیر درمی‌آید:

$$\min_{w,\xi_i} \tfrac{۱}{۲}\|W\|_۲^۲ + C\sum_{i=۱}^{n}\xi_i$$

$$\text{s.t. } y_i(W^T(x_i)) \geq ۱ - \xi_i, \forall i \in \{۱, \cdots n\}$$

$$\xi_i \geq ۰, \forall i \in \{۱, \cdots, n\}$$

که $\xi_۱, \cdots, \xi_n$ متغیر های slack نامیده می‌شوند. فرض کنید $\xi_۱, \cdots, \xi_n$ بهینه محاسبه شده‌اند. با استفاده از ξ_i یک کران بالا برای تعداد نمونه‌هایی که به غلط دسته‌بندی شده‌اند بیابید.

۳. (۲ نمره) در بهینه سازی اصلی (primal) SVM نقش C چیست؟ پاسخ خود را با توجه به دو حالت $C \to ۰$ و $C \to \infty$ به طور خلاصه شرح دهید.

۴. (۲ نمره) در حالتی که دو کلاس خطی تفکیک پذیر هستند Hard SVM و Logistic Regression را مقایسه کنید. هر تفاوت عمده را بیان کنید. (راهنمایی: به مرز تصمیم فکر کنید.)

۵. (۲ نمره) در حالتی که دو کلاس خطی تفکیک پذیر نیستند Soft SVM و Logistic Regression را مقایسه کنید. هر تفاوت عمده را بیان کنید.

۲ قسمت عملی

۱.۲ Soft Margin Support Vector Machine (۳۰ + ۱۰ نمره)

مجموعه دادگان satimage را دانلود کنید. این مجموعه داده که شش کلاس است، از پیش به سه قسمت Train ، Validation و Test تقسیم شده است و با فرمت خاص پکیج LIBSVM ذخیره شده که برای خواندن آن می توانید از دستور

sklearn.datasets.load-svmlight-file('filename')

استفاده کنید.

۱. (۱۲ نمره) soft margin را با $C = ۱$ برای تمایز بین دو دسته‌ی ۴ و ۶ پیاده‌سازی کنید. برای این منظور صورت دوگان مسئله را با استفاده از پکیج‌های quadratic solver مثل cvxopt حل کنید. دقت را روی دادگان تست گزارش کنید. همچنین confusion matrix و balanced accuracy را گزارش کنید. (در تسک دسته بندی k کلاسه، confusion matrix یک ماتریس $k \times k$ است که خانه‌ی $[i, j]$ آن، تعداد نمونه‌هایی را نشان می دهد که در اصل عضو دسته‌ی i

[1] Decision Boundary

بودند و دسته‌بند آن‌ها را عضو دسته‌ی j پیش‌بینی کرده‌است. و اما در مورد balanced accuracy : این معیار برخلاف accuracy نتیجه‌اش متاثر از تعداد زیاد اعضای یک دسته نیست. فرض کنید در همین سوال، ۹۵ درصد داده‌های تست متعلق به کلاس ۶ بودند. در این صورت دسته‌بندی که کلا همه‌ی داده‌ها را کلاس ۶ ام پیش‌بینی کند دقت خوب ۹۵ درصد را کسب می‌کند اگرچه اصلا دسته‌بند مطلوبی نیست. برای جبران این ضعف، balanced accuracy به صورت میانگین accuracy روی داده‌های هرکلاس تعریف می‌شود. لذا فرمول آن به صورت $\frac{\sum_{i=1}^{k} ACC_i}{k}$ است که ACC_i میزان دقت دسته‌بند روی داده‌های فقط دسته i ام است.

۲. (۱۲ نمره) این بار قسمت قبل را با کرنل rbf انجام دهید. برای تعیین مقدار مناسب هایپرپارامتر σ از مجموعه دادگان اعتبارسنجی استفاده کنید. دقت و دقت متعادل دسته‌بند حاصل را روی داده‌ی تست گزارش کنید.

۳. (۶ نمره) در این قسمت بدون پیاده‌سازی و با استفاده از تابع آماده SVC از پکیج sklearn دسته‌بند soft margin را روی کل مجموعه دادگان یادگیری(کل شش کلاس) اجرا کنید. هایپرپارامترها و آرگومان‌های تابع را تغییر دهید و با بررسی نتایجشان رو دادگان اعتبارسنجی، عملکرد دسته‌بند را تا جای ممکن بهبود بخشید. معیار سنجشتان برای مدل‌ها accuracy باشد. در آخر عملکرد مدل نهایی را روی دادگان آزمایش گزارش کنید. در ضمن، روندی که به این انتخاب خاص از آرگومان‌های تابع منجر شد را با گزارش عملکرد مدل‌های میانی روی دادگان اعتبارسنجی توضیح دهید.

۴. ((۱۰ نمره)امتیازی) دسته‌بند چند کلاسه (حداقل ۳ کلاسه) soft margin را پیاده‌سازی کنید و دقتش را روی دادگان تست گزارش کنید.

موفق باشید