یادگیری ماشین

نیم‌سال اول ۹۹ـ۱۴۰۰

دانشکده‌ی مهندسی کامپیوتر

مدرس: دکتر حسینی

تمرین سری ششم

موعد تحویل:۲۵ آذر ۹۹

۸۷ نمره نظری + ؟ نمره عملی

مسئله‌ی ۱. ماشین بردار پشتیبان (۲۶ نمره)

الف) (۱۰ نمره) در این سوال هدف مرور ماشین بردار پشتیبانی با حاشیه نرم با تابع هزینه زیر است:

$$\min_{w_i, w_0, \xi_i} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N} \xi_i$$

$$s.t. \quad y^{(i)}(w^T x(i) + w_0) \geqslant 1 - \xi_i \quad , \quad \xi_i \geqslant 0 \ i = 1, ..., N$$

یک) (۲ نمره) اگر مجموعه داده‌ها به صورت خطی جداپذیر باشند، آیا جواب هر دو نسخه نرم و سخت SVM حتما یکی خواهد بود؟ در صورتی که جواب بله است چرا و در صورتی که جواب خیر است توضیح دهید با چه فرض در صورت مساله بهینه‌سازی این اتفاق رخ خواهد داد.

دو) (۲ نمره) هر یک از حالات $\xi_i > ۱$ و $۱ > \xi_i > ۰$ و $\xi_i = ۰$ و $\xi_i = ۱$ از نظر توصیفی به چه معنا هستند؟

سه) (۲ نمره) فرم‌های اولیه و دوگان SVM را با یکدیگر مقایسه کنید و توضیح دهید که به چه دلیل استفاده از فرم دوگان مرجح است؟

چهار) (۴ نمره) در یک نمودار تابع ضرر SVM با حاشیه‌های سخت و نرم، پرسپترون و لاجیستیک رگرشن را برای وقتی که برچسب مطلوب نمونه یک است را رسم نموده و با یکدیگر مقایسه نمایید.

ب) (۴ نمره) فرض کنید به جای استفاده از تابع هزینه رایج برای SVM با حاشیه نرم، از تابع هزینه زیر استفاده شود.

$$\min_{w_i, w_0, \xi_i} \frac{1}{2}||w||^2 + C\sum_{i=1}^{N} \xi_i^2$$

با استفاده از روابط ریاضی تحلیل کنید که مدل به دست آمده از این تابع هزینه، چه تفاوتی در میزان حساسیت به داده‌های نویز، با مدل به دست آمده از تابع هزینه عادی دارد.

پ) (۱۲ نمره) می‌خواهیم با استفاده از SVM حاشیه سخت و با تابع هزینه به فرم دوگان آن $x_۱$ تا $x_۵$ را دسته‌بندی کنیم. برچسب داده‌های $x_۱$ و $x_۲$ برابر $+۱$ و برچسب باقی داده‌ها برابر با -۱ است. فرض کنید برای داده‌ها داریم:

$$D = \{x_۱ = (۰, ۱), x_۲ = (۱, ۰), x_۳ = (۱, ۲), x_۴ = (۲, ۲), x_۵ = (۱, ۳)\}$$

یک) (۲ نمره) رابطه بهینه‌سازی (فرم دوگان) را نوشته و مقادیر عددی مربوط به این مساله را در آن جایگذاری کنید.

دو) (۲ نمره) نقاط داده شده را در فضای ۲ بعدی رسم نمایید و مرز تصمیم‌گیری را نیز به صورت تقریبی بر روی نمودار رسم کنید. نقاطی که بردار پشتیبان نمی‌توانند باشند را نیز مشخص کنید.

سه) (۲ نمره) با استفاده از نتیجه‌ی قسمت قبل، رابطه‌ی قسمت قبل یک را ساده نمایید.

چهار) (۳ نمره) حال بدون حل مستقیم مسئله QP و با کمک رابطه به دست‌آمده در قسمت قبل، سعی کنید مسئله بهینه‌سازی را حل نمایید.

پنج) (۳ نمره) رابطه‌ی مربوط به مرز تصمیم‌گیری را به دست آورید.

مسئله‌ی ۲. هسته (۲۱ نمره)

الف) (۸ نمره) فرض کنید که $k_۱(x,x')$ و $k_۲(x,x')$ دو هسته معتبر باشند. در این صورت معتبربودن هسته‌های زیر را نیز اثبات نمایید.

یک) (۲ نمره) $k_۳(x,x') = k_۱(x,x') + k_۲(x,x')$

دو) (۲ نمره) $k_۴(x,x') = k_۱(x,x').k_۲(x,x')$

سه) (۱ نمره) $k_۵(x,x') = ak_۱(x,x')$) به شرطی که $a \geqslant ۰$

چهار) (۳ نمره) $k_۵(x,x') = exp(k_۱(x,x'))$)

ب) (۳ نمره) فرض کنید A و B زیرمجموعه‌هایی از مجموعه مرجع S هستند. ثابت کنید هسته $k(A,B) = ۲^{|A\cap B|}$ برای داده‌ساختار مجموعه معتبر است.

پ) (۱۰ نمره) هسته چندجمله‌ای به شکل $k(x,x') = (x^T.x' + c)^M$ را فرض کنید که در آن داریم $x, x' \in R^d$ و $c \geqslant ۰$

یک) (۳ نمره) اگر $M = ۲$ باشد، به این هسته، هسته‌ی درجه‌ی دوم می‌گویند. تابع نگاشت ویژگی $\phi(x)$ مرتبط با هسته‌ی درجه دوم را به دست آورید.

دو) (۲ نمره) اگر $c = ۰$ در نظر گرفته شود، فضای تبدیل چه تغییری می‌کند؟

سه) (۵ نمره) تعداد ابعاد فضای ویژگی جدید را برای حالت کلی چندجمله‌ای درجه M و $c > ۰$ به دست آورید.

مسئله‌ی ۳. درخت تصمیم (۲۰ نمره)

فرض کنید N داده‌ی آموزش به صورت $\mathcal{D} = \{(x^{(۱)}, y^{(۱)}), ...(x^{(N)}, y^{(N)})\}$ داریم که $x^{(n)} \in \mathcal{R}^D$ و $y^{(n)} \in \{۱, ..., C\}$. مولفه‌ی d ام از $x^{(n)}$ را با $x_d^{(n)}$ نشان می‌دهیم و آن را ویژگی d ام از نمونه‌ی n ام می‌نامیم. در واقع از هر نمونه D تا ویژگی داریم که می‌توانند حقیقی باشند. هر نمونه به یکی از C کلاس تعلق دارد.

الف) (۶ نمره) نشان دهید اگر ویژگی d ام و y از هم مستقل باشند آنگاه $Gain(y, x_d) = ۰$. یعنی الگوریتم ID3 ویژگی d ام را به عنوان ریشه‌ی درخت تصمیم انتخاب نمی‌کند. (فرض‌کنید در داده‌ها نویز وجود ندارد و به اندازه‌ی کافی نمونه داریم)

ب) (۶ نمره) فرض‌کنید N نفر داریم که از هر کدام D ویژگی داریم. اگر بیماری فرد n ام خوش‌خیم بود $y^{(n)} = ۰$ و در غیر این صورت $y^{(n)} = ۱$. فرض کنید $x_۱^{(n)}$ شماره‌ی پرونده‌ی بیمار n ام است. همچنین می‌دانیم شماره‌ی پرونده منحصر به فرد است. اگر برای حل این مسئله از الگوریتم ID3 و معیار Information Gain استفاده کنیم، آیا ممکن است ویژگی اول به عنوان ریشه‌ی درخت تصمیم قرار گیرد؟ اگر این اتفاق بیفتد، آیا Overfitting رخ داده است؟ اگر پاسخ مثبت است راهکاری برای جلوگیری از این نوع Overfitting ارائه دهید.

پ) (۸ نمره) تعریف(درخت تصمیم بدون تکرار): فرض کنید یک درخت تصمیم داریم که در هر مسیر مستقیم از ریشه به یک برگ، هر ویژگی حداکثر یک‌بار بررسی می‌شود. چنین درختی را بدون تکرار می‌نامیم.

یک) (۴ نمره) فرض کنید ویژگی‌ها مقادیر گسسته و محدودی دارند (مثلا $x_d^{(n)} \in \{۰, ۱, ..., \alpha_d\}$). نشان دهید درخت تصمیم بدون تکراری وجود دارد که خطای آموزش آن صفر است.

دو) (۲ نمره) فرض کنید ویژگی‌ها پیوسته هستند. همچنین وقتی در یک گره‌ی درخت یک ویژگی پیوسته را بررسی می‌کنیم، روی آن آستانه می‌گذاریم. اگر ویژگی از آستانه کمتر بود به زیردرخت چپ و اگر بیشتر بود به زیردرخت راست می‌رویم. آیا به ازای هر داده‌ی آموزش با ویژگی‌های پیوسته، درخت تصمیمی وجود دارد که ۱. به شکل فوق ویژگی‌های پیوسته را بررسی کند و ۲. بدون تکرار باشد و ۳. خطای آموزش آن صفر باشد؟

سه) (۲ نمره) الگوریتم ID3 در هر مرحله یک ویژگی را به عنوان ریشه انتخاب می‌کند و در آن زیردرخت، دیگر از آن ویژگی استفاده نمی‌کند. با توجه به قسمت قبل چرا در حالت پیوسته (برخلاف حالت گسسته) نباید ویژگی ریشه را کنار گذاشت؟

مسئله‌ی ۴. یادگیری جمعی (۱۲ نمره)

فرض کنید در یک مسئله‌ی رگرسیون، تابع مطلوب $h(x)$ باشد. M تابع به نام‌های $h_۱(x), h_۲(x), ..., h_M(x)$ در نظر بگیرید. تابع ترکیبی $H_M(x)$ را به صورت $H_M(x) = \frac{1}{M}\sum_{m=1}^{M} h_m(x)$ در نظر می‌گیریم. امید ریاضی مجذور خطا برای $H_M(x)$ روی کل توزیع را E_{com} می‌نامیم. همچنین میانگین امید ریاضی مجذورات خطا برای توابع $h_۱(x), h_۲(x), ..., h_M(x)$ را E_{avg} می‌نامیم. به عبارت دیگر، E_{com} و E_{avg} از روابط زیر به دست می‌آیند:

$$E_{com} = E_x[(H_M(x) - h(x))^2]$$

$$E_{avg} = \frac{1}{M}\sum_{m=1}^{M} E_x[(h_m(x) - h(x))^2]$$

الف) (۶ نمره) نشان دهید $E_{com} \leqslant E_{avg}$.

(راهنمایی: می‌توانید از این موضوع استفاده کنید که اگر تابع f محدب باشد، آنگاه $f(E[x]) \leqslant E[f(X)]$)

ب) (۶ نمره) اگر امید خطای مدل‌های پایه ناهمبسته باشند،

$$\forall\ m \neq l \quad E[(h_m(x) - h(x))(h_l(x) - h(x))] = 0$$

نشان دهید $E_{com} = \frac{1}{M}E_{avg}$.

مسئله‌ی ۵. Adaboost (۸ نمره)

الف) (۴ نمره) فرضیه به دست آمده در دو مرحله‌ی متوالی از Adaboost را h_t و h_{t+1} می‌نامیم. با فرض این که خطای یادگیرهای ضعیف از 0.5 اکیدا کمتر است، نشان دهید $h_t \neq h_{t+1}$.

ب) (۴ نمره) نشان دهید در الگوریتم Adaboost می‌توان وزن‌ها را طوری به‌روزرسانی کرد که ۱.فقط وزن داده‌های به اشتباه دسته‌بندی شده تغییر کند. ۲. نتایج الگوریتم تغییر نکند.

مسئله‌ی ۶. (عملی) پیاده‌سازی SVM (۲۵ نمره)

دادگان Digits را که از این طریق قابل بارگیری است، در نظر بگیرد.

این دادگان شامل تعدادی تصاویر از ارقام دست‌نویس زبان انگلیسی است. به ازای هر تصویر یک بردار با ۶۴ درایه وجود دارد و این درایه‌ها میزان روشنایی پیکسل‌های یک تصویر ۸ در ۸ را نشان می‌دهند. با توجه به این که برچسب هر یک از تصاویر یکی از ارقام ۰ تا ۹ می‌تواند باشد ، در این جا قصد داریم تا با روش One against All این تصاویر را دسته‌بندی نماییم. به این منظور شما بایستی روش SVM با دو حالت هسته خطی و هسته‌گاوسی را طبق موارد خواسته شده در زیر و با استفاده از بهینه‌سازی QP پیاده‌سازی کنید. دقت کنید که در این تمرین نمی‌توانید از توابع حاضر و آماده برای آموزش مدل SVM استفاده کنید و تنها مجاز به استفاده از کتابخانه CVXOPT می‌باشید.

الف) (۵ نمره) ابتدا دادگان را با نسبت ۹۰ و ۱۰ به دو بخش آموزش و تست تقسیم کنید. اطمینان حاصل کنید که از هر یک از برچسب‌ها حداقل ۱۵ داده در مجموعه دادگان تست حاضر باشد. سپس مقدار هر یکی از ویژگی‌ها را به مقداری بین صفر تا یک به روشی دلخواه نرمالایز کنید.

ب) (۱۰ نمره) SVM با حاشیه نرم را پیاده‌سازی کرده و روی دادگان آموزش دهید.
برای انتخاب مقدار بهینه پارامتر C از اعتبارسنجی 4-fold CV استفاده کنید.
پس از آموزش دقت مدل بهینه را روی مجموعه دادگان آموزش و تست گزارش کنید.

پ) (۱۰ نمره) حال SVM با حاشیه نرم با هسته گاوسی را روی دادگان آموزش دهید.
برای انتخاب ابرپارامتر در این جا نیز مانند قسمت قبل عمل کنید.
پس از آموزش دقت مدل بهینه را روی مجموعه دادگان آموزش و تست گزارش کنید.

موفق باشید