حل تمرینات شناسایی الگو - سوال ۹

وحید ملکی

۲۲ آذر ۱۴۰۴

سوال ۹ - بخش الف: تاثیر پارامتر تنظیم C در SVM

در ماشین بردار پشتیبان (SVM)، پارامتر C نقش جریمه خطاها را بازی می‌کند و تعادلی میان «ماکزیمم کردن حاشیه» (Margin) و «کمینه کردن خطای دسته‌بندی» ایجاد می‌کند.

۱. شکل سمت چپ ($C = 0.05$)

تحلیل: مقدار C بسیار کوچک است. این یعنی جریمه خطا کم است و مدل ترجیح می‌دهد حاشیه (Margin) را پهن‌تر نگه دارد، حتی اگر تعدادی از داده‌ها اشتباه دسته‌بندی شوند یا درون حاشیه قرار بگیرند (Soft Margin).

مرز تصمیم گیری: یک خط با جهت‌گیری کلی که سعی دارد توده اصلی داده‌ها را جدا کند، اما نسبت به داده‌های پرت (Outliers) حساسیت کمی دارد.

بردارهای پشتیبان (Support Vectors): در حالت حاشیه نرم، هر داده‌ای که:

۱. دقیقاً روی مرز حاشیه باشد،

۲. درون ناحیه حاشیه (بین دو خط‌چین فرضی) قرار داشته باشد،

۳. یا در سمت اشتباه دسته‌بندی شده باشد،

یک بردار پشتیبان است. بنابراین در شکل سمت چپ، تعداد بردارهای پشتیبان **زیاد** است (شامل داده‌های پرت دایره و ستاره و همچنین داده‌های نزدیک به خط مرزی).

۲. شکل سمت راست ($C = 10000$)

تحلیل: مقدار C بسیار بزرگ است. این یعنی جریمه خطا بسیار سنگین است و مدل تمام تلاش خود را می‌کند تا تمام داده‌ها را درست دسته‌بندی کند (رفتار شبیه به Hard Margin). مدل حاشیه را باریک می‌کند تا بتواند از بین داده‌های نزدیک به هم عبور کند.

مرز تصمیم گیری: خطی که سعی کرده خود را با داده‌های مرزی دقیقاً تنظیم کند تا خطایی رخ ندهد.

بردارهای پشتیبان: در این حالت حاشیه بسیار باریک است. تنها داده‌هایی که دقیقاً مماس بر حاشیه هستند (نزدیک‌ترین داده‌ها به خط جداکننده) به عنوان بردار پشتیبان انتخاب می‌شوند. بنابراین در شکل سمت راست، تعداد بردارهای پشتیبان **کم** است (احتمالاً تنها ۱ یا ۲ داده از هر کلاس که کمترین فاصله را تا خط دارند).

سوال ۹ - بخش ب: نظیر کردن کرنل‌ها و شکل‌ها

در این بخش باید توصیف‌های (الف) تا (و) را به شکل‌های ۱ تا ۶ نظیر کنیم.

- **الف) دسته‌بند SVM خطی با حاشیه تمایز نرم و $C = 0.1$:**
  این گزینه متناظر با **شکل ۳** است.
  دلیل: مرز خطی است. چون C کوچک است، حاشیه (فاصله خط‌چین‌ها تا خط ممتد) پهن و عریض در نظر گرفته شده است و برخی داده‌ها به درون حاشیه نفوذ کرده‌اند.

- **ب) دسته‌بند SVM خطی با حاشیه تمایز نرم و $C = 10$:**
  این گزینه متناظر با **شکل ۴** است.
  دلیل: مرز خطی است. چون C بزرگتر است، حاشیه باریک‌تر شده و سخت‌گیری بیشتری روی داده‌ها اعمال شده است (خط‌چین‌ها به خط ممتد نزدیک‌ترند).

- **ج) دسته‌بند SVM غیرخطی با کرنل $K(u, v) = u^t v + (u^t v)^2$:**
  این گزینه متناظر با **شکل ۱** است.
  دلیل: این یک کرنل چندجمله‌ای درجه ۲ است. مرزهای تصمیم‌گیری حاصل از درجه ۲ می‌توانند مقاطع مخروطی (بیضی، دایره، سهمی و ...) باشند. شکل ۱ یک مرز بیضی‌شکل بسته را نشان می‌دهد که مشخصه بارز کرنل‌های درجه ۲ است.

- **د) دسته‌بند SVM غیرخطی با کرنل گاوسی و $\gamma = 0.25$:**
  این گزینه متناظر با **شکل ۵** است.
  دلیل: فرمول کرنل گاوسی $exp(-\gamma||u - v||^2)$ است. در اینجا $\gamma = 1/4 = 0.25$ مقدار کوچکی است. گامای کوچک باعث ایجاد واریانس بالا و مرزهای تصمیم‌گیری **هموار و نرم (Smooth)** می‌شود که کلیت داده‌ها را جدا می‌کند. شکل ۵ یک منحنی هموار است.

- **هـ) دسته‌بند SVM غیرخطی با کرنل گاوسی و $\gamma = 4$:**
  این گزینه متناظر با **شکل ۶** است.
  دلیل: در اینجا $\gamma = 4$ مقدار بزرگی است. گامای بزرگ باعث می‌شود مدل بسیار موضعی عمل کند و دور تک‌تک داده‌ها یا دسته‌های کوچک «جزیره» ایجاد کند (Overfitting). شکل ۶ دقیقا نشان‌دهنده این رفتار است که یک مرز بسته‌ی جزیره‌ای دور گروهی از داده‌ها تشکیل شده است.

نکته: شکل ۲ در گزینه‌ها استفاده نشد (یا می‌تواند حالت دیگری از کرنل چندجمله‌ای باشد، اما شکل ۱ برای چندجمله‌ای و شکل ۵ و ۶ برای گاوسی گزینه‌های دقیق‌تری هستند).