باسمه تعالی

دانشگاه صنعتی شریف

دانشکده مهندسی برق

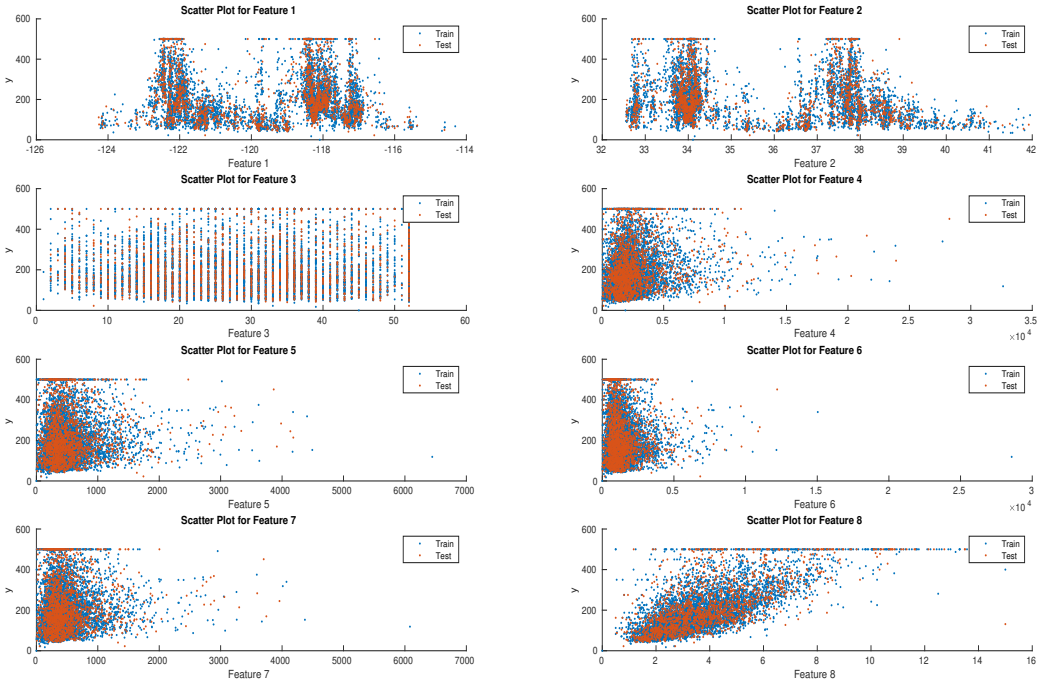
مقدمه‌ای بر یادگیری ماشین ــ دکتر جمال‌الدین گلستانی

بهراد منیری ــ ۹۵۱۰۹۵۶۴

گزارش بخش کامپیوتری تمرین سری اول

۱ تمرین اول

شکل زیر، نمودارهای Scatter Plot داده‌های این تمرین، برای هر ویژگی هستند. داده‌های تست ، با رنگ قرمز و داده‌های آموزشی، با رنگ آبی مشخص شده‌اند.



۱.۱ بخش الف

ویژگی X برای رگرسیون مناسب است اگر X و y مستقل نباشند. برای سنجش استقلال متغیرهای تصادفی، می‌توان از آزمون فرضیه‌ی Hilbert-Schmidt استفاده کرد. با توجه به نمودارهای، به صورت شهودی، به نظر می‌رسد که ویژگی شماره‌ی هشت بیشترین وابستگی را به y داشته باشد و در نتیجه مناسب‌ترین انتخاب برای عمل رگرسیون است. بدترین ویژگی نیز ویژگی شماره‌ی ۳ است که به نظر مقدار آن کاملاً مستقل از y است.

۲.۱ بخش ب

در متلب، به کمک نتایجی که در کلاس برای کمینه‌کردن MSE به دست آمد، رگرسیون خطی انجام می‌دهیم و خطای این رگرسیون را، یک بار بر روی خود داده‌های آموزشی، و یک بار بر روی داده‌های کنارگذاشته‌شده در یادگیری محاسبه می‌کنیم. در این‌جا نیز از Squared Error برای خطا استفاده می‌کنیم. مطابق انتظار، خطا بر روی داده‌های آموزشی، کم از از خطا بر روی داده‌های تست به دست آمد.

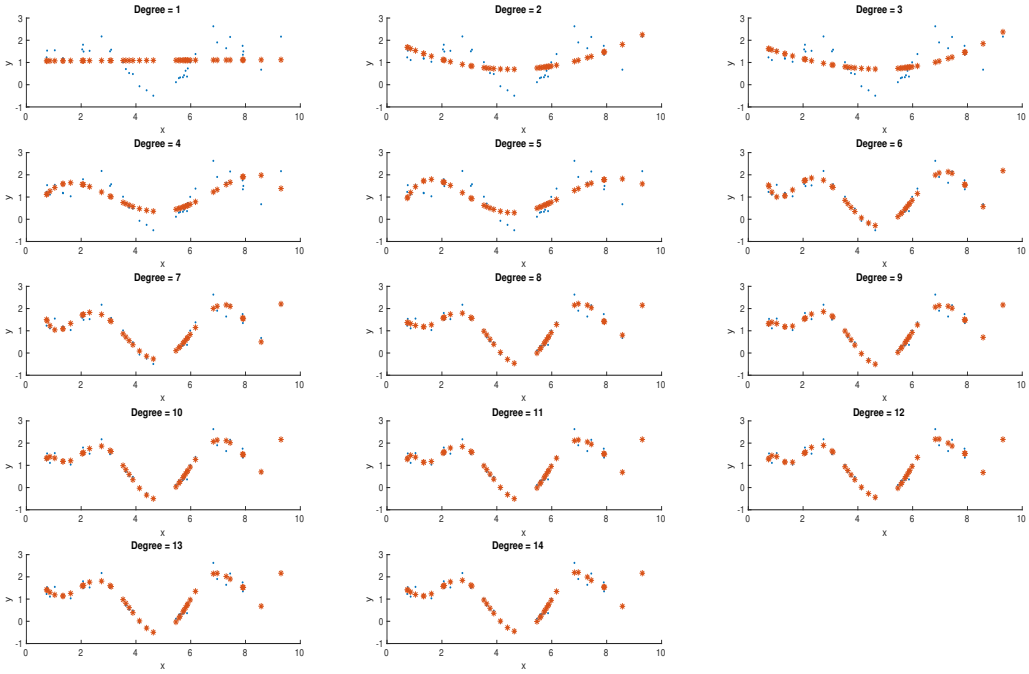
مقدار تلف، برای داده‌های تست، True Risk، برابر 5.598×10^3 و برای داده‌های آموزشی، Empirical Risk، 4.726×10^3 به دست آمد.

بردار ضرایب نیز برابر

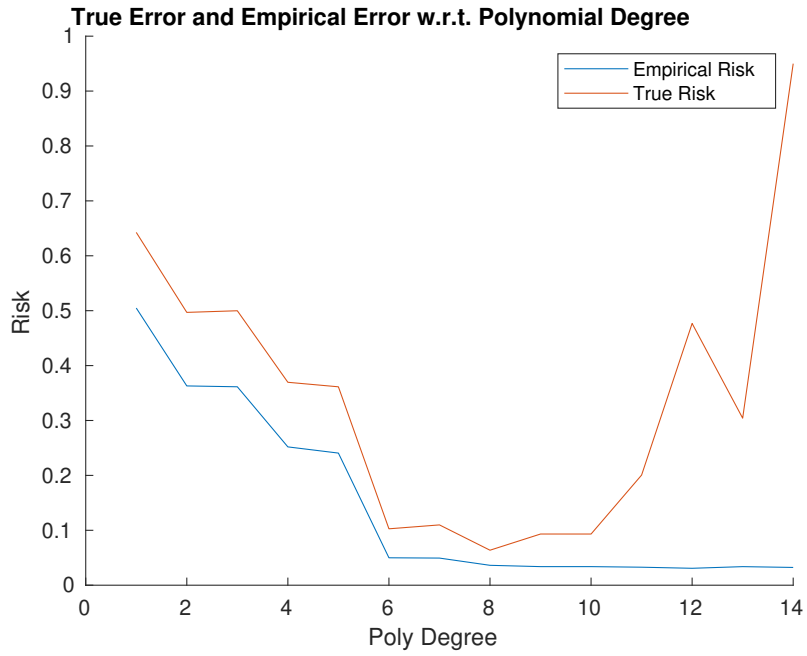
$$\omega = 10^3 \times [-0.04169, -0.04186, 0.00117, -0.00001, 0.00010, -0.00004, 0.00006, 0.03896, -3.48358]$$

۲ تمرین دوم

داده‌ها را در محیط دوبعدی رسم می‌کنیم و آنها را با چندجمله‌ای هایی از درجات مختلف رگرس می‌کنیم. نقاط آبی، داده‌های و نقاط قرمز، پیش‌بینی ما برای داده‌های آموزشی است.



نمودار Empirical Risk و True Risk بر حسب درجات مختلف چندجمله‌ای، به صورت زیر است.



خطای واقعی به ازای چندجمله‌ای درجه‌ی هشت کیمنه می‌شود.

همان‌طور که در درس دیدیم، با زیاد کردن درجه، کلاس فرضیه‌های خود را بزرگتر می کنیم و در نتیجه خطا بر روی داده‌های آموشی کمتر می‌شود یا ثابت می‌ماند. این اتفاق به وضوح در نمودار دیده می‌شود. خطای واقعی نیز با غنی کردن مجموعه‌ی فرضیه‌های کاهش می‌یابد ولی بزرگ‌کردن بیش از حد آن باعث می‌شود تا دچار Overfitting شویم.