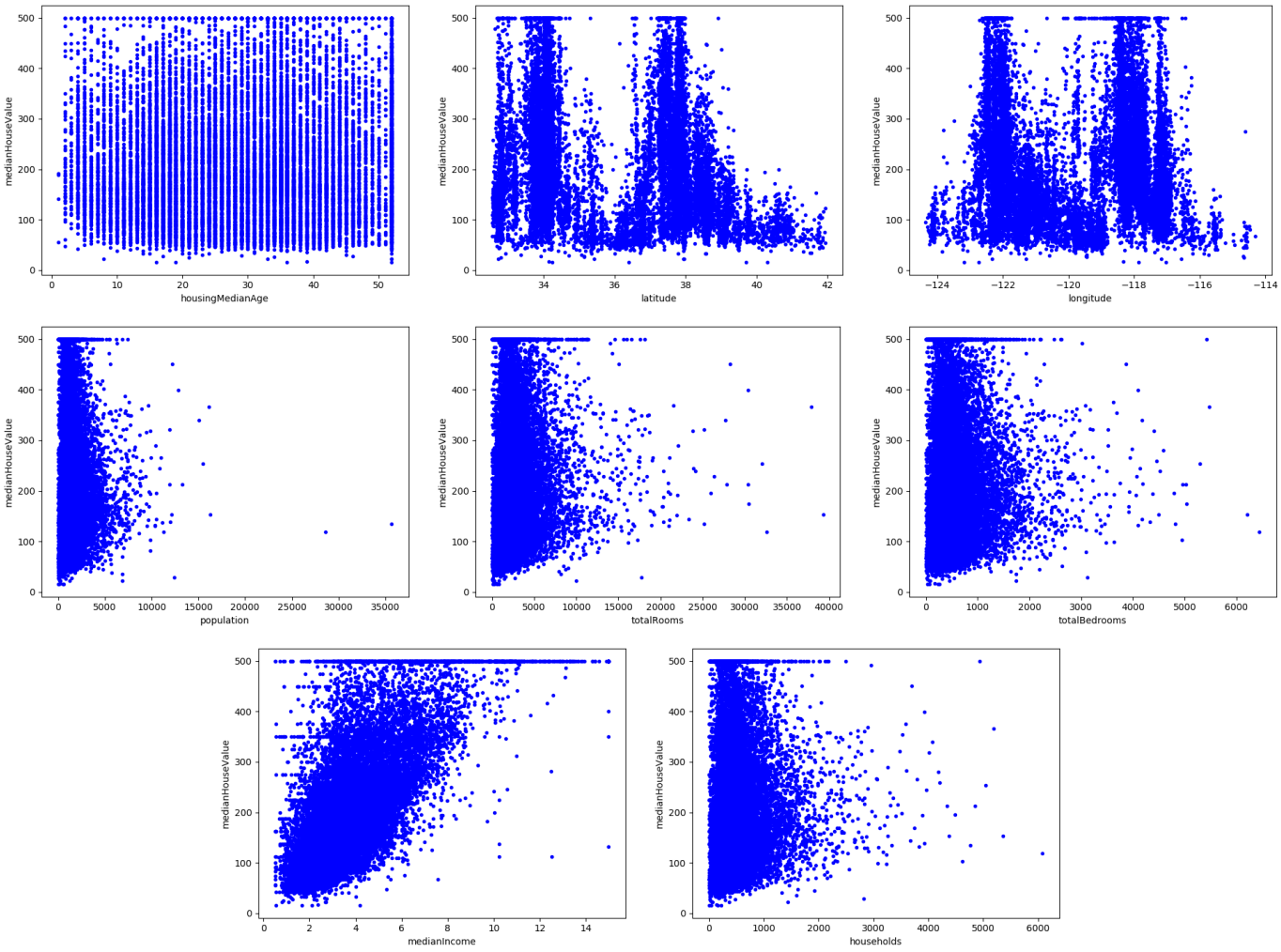
به نام خدا

گزارش سوالات عملی تمرین اول
مقدمه‌ای بر یادگیری ماشین

حسین ابراهیمی ـ ۹۵۱۰۵۳۰۲

مدرس : دکتر جمال‌الدین گلستانی

سوال C۱



شکل ۱: نمودار قیمت بر حسب ویژگی‌های مختلف

- بهترین ویژگی برای پیش‌بینی قیمت medianIncome است زیرا چون تقریبا یک رابطه خطی با medianHouseValue دارد و به ازای هر medianIncome ثابت بازه‌ای قابل اعتبار (valid interval) کوچکتری نسبت به سایر ویژگی‌ها دارد. از لحاظ منطقی نیز این رابطه درست است چون هر چه افراد دارای درآمد بالاتری باشند، قیمت خانه‌ی آن‌ها نیز به نسبت بالاتر می‌رود و همچنین می‌شود با دانستن قیمت خانه‌ی هر فرد به طور تقریبی متوسط درآمد او را فهمید به همین دلیل این رابطه‌ی خطی بین این دو کاملا با PriorKnowledge ما همخوانی دارد و برای پیش‌بینی از سایر ویژگی‌ها مناسب‌تر است.
- ابتدا ۲۵ درصد انتهایی داده را برای validation جدا می‌کنیم سپس قیمت خانه که همان Y است، بر حسب ستون‌های X رسم می‌کنیم که همان نمودارهایی است که در بالا می‌بینیم. سپس برای گذر از مبدا یک ستون ۱ به داده X خود اضافه می‌کنیم و با استفاده از رابطه

$$w^* = A^{-1}b = (X^T X)^{-1} X^T Y$$

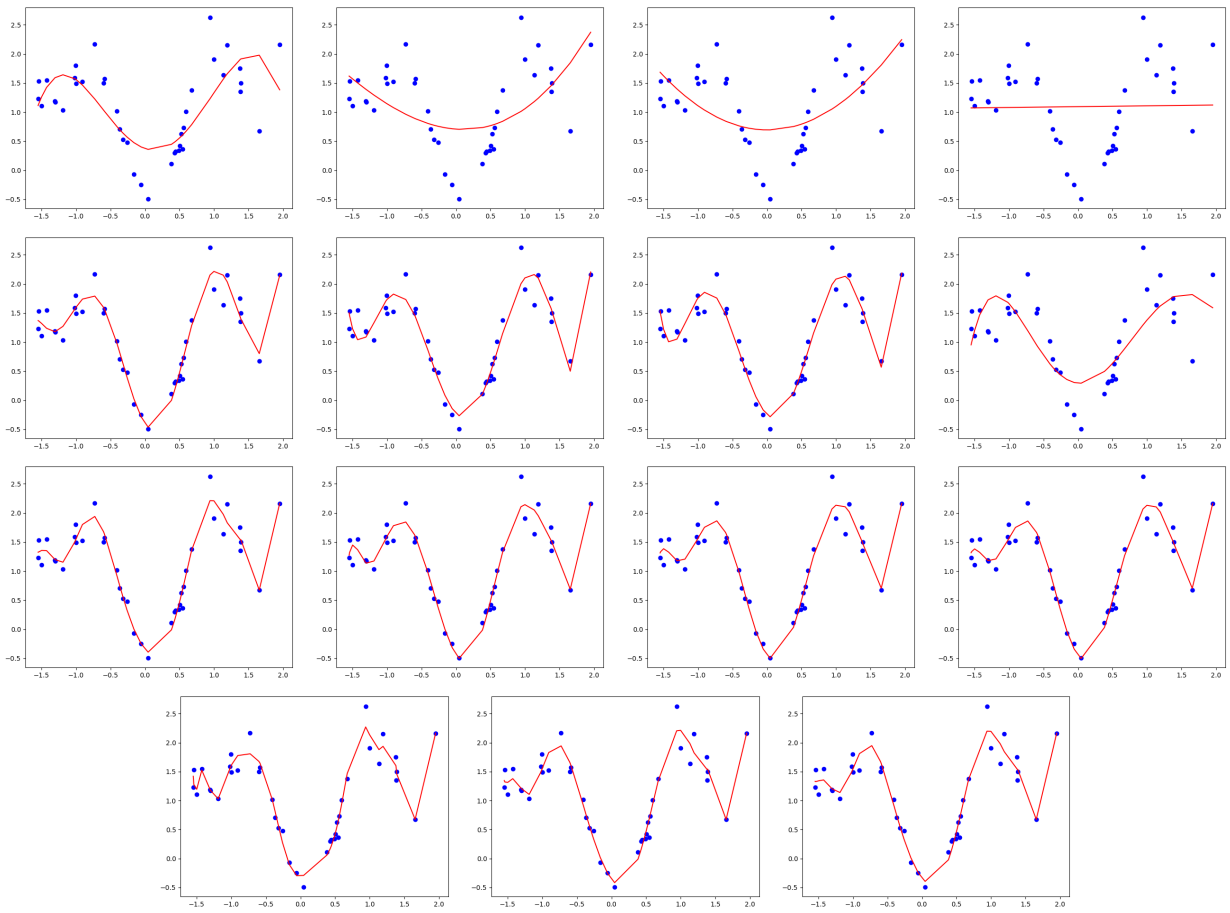
بهترین ضرایب ممکن را بدست می‌آوریم که برابر است با :

$$weights = [-3570.33836457,\ -42.67465338,\ -42.7079347,\ 1.19272055,$$
$$-0.00651816,\ 0.10282147,\ -0.04326593,\ 0.06300437,\ 39.75544804]$$

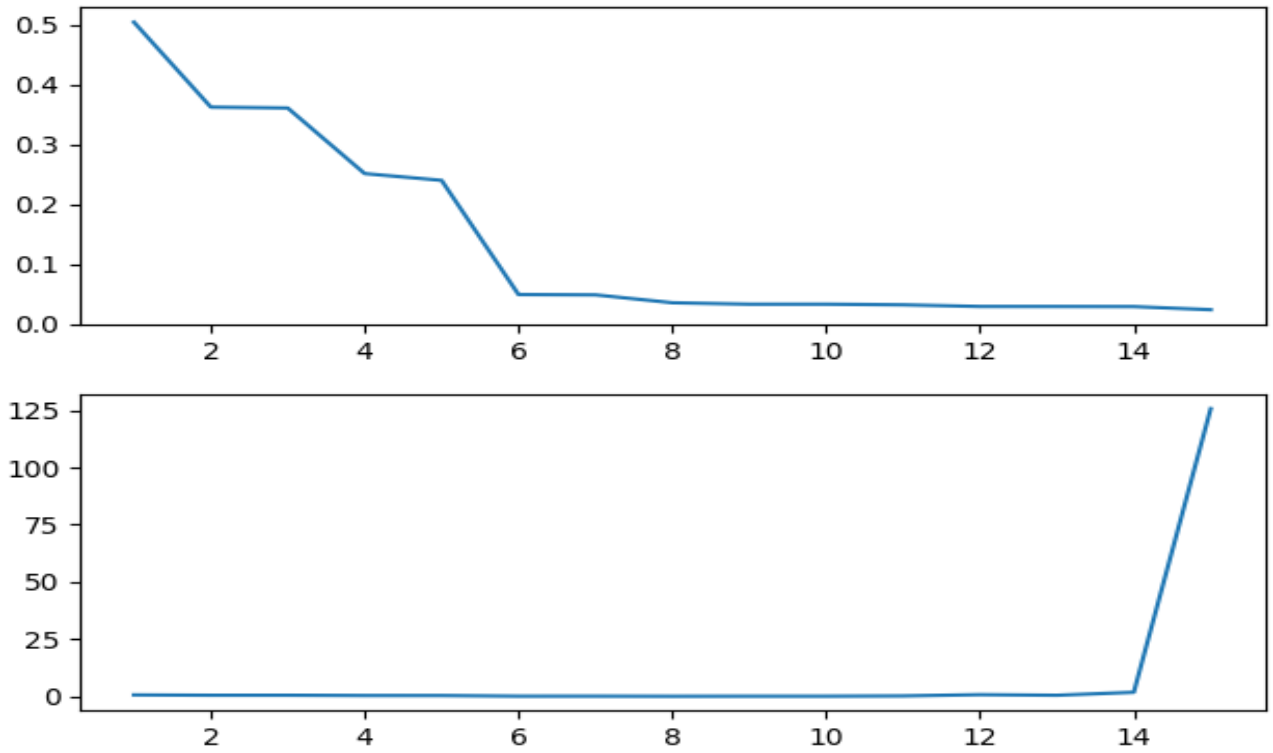
حال با استفاده از رابطه $loss = \frac{1}{m}\sum_{i=1}^{m}(w^T x_i - y_i)^2$، empirical risk و true risk را بدست می‌آوریم که به شکل زیر است:

$$empirical\ risk = 4780.70077233673$$
$$true\ risk = 5007.513797793691$$

سوال C۲



شکل ۲: نمودارهای رگرسیون چندجمله‌هایی از درجه ۱ تا ۱۵ از راست به چپ



شکل ۳: خطای تجربی در نمودار بالا و خطای واقعی در نمودار پایین بر حسب درجه‌ی چندجمله‌ای

- برای محاسبه بهترین رگرسیون چند جمله‌ای برای درجات بالاتر از ۹ در پایتون به مشکل ($overflow$) اشتباه بودن اعدادی که از توان‌های بالای داده می‌آیند برخوردم و برای رفع این مشکل کل X داده‌ها اعم از آموزشی و تست را در بازه بین $[-1, 1]$ نرمال سازی کردم به این صورت که داده‌ی جدید با استفاده رابطه‌ی زیر بدست آمد:

$$x'_i = \frac{x_i - \mu}{s.t.d} \qquad \mu = Average\ of\ Data, \quad s.t.d = Standard\ Deviation$$

حال با استفاده از داده‌ی جدید که مشکلی در توان‌های بالا ندارد، ضرایب هر رگسیون چندجمله‌ای و همچنین خطای آموزشی و خطای واقعی را برای هر کدام بدست آوردم. نمودارهای بالا نیز بر حسب داده‌ی جدیدی که نرمال شده است بدست آمده.

$$empirical\ risk = [0.5048439684237046, 0.36296074571951153, 0.36140360499600793, 0.2519239212995267, 0.2406392617857301, 0.04993652294947328, 0.049397470426849006, 0.03622579947869632, 0.033840774320160534, 0.033837344887134045, 0.032782850732900756, 0.029998331632671467, 0.02994581424262386, 0.029782762708386562, 0.024618051124705453]$$

- خطای واقعی بر حسب چند جمله با درجه ۸ کمترین خطا را داراست.

$$true\ risk = [0.6425763922624383, 0.49690283695592063, 0.49987278075469244, 0.3695918303856628, 0.36136078514160735, 0.10271625441528556, 0.10982257180990965, 0.06373926993481935, 0.09321857814401026, 0.09319868907496402, 0.2005744021766423, 0.7473482642390301, 0.48813984706838864, 1.795506788427145, 125.73648125314173]$$

- علت تفاوت بین تعییرات خطای واقعی و خطای تجربی آن است که با بالا بردن درجه‌ی چندجمله‌ای، مدل ما هر چه بیشتر شبیه به داده‌ی آموزشی ما می‌شود و خطای آموزشی ما کاهش می‌یابد و این انتظار را داریم که این رابطه همواره روی داده تست ما هم صادق باشد و خطای واقعی پیوسته کاهش پیدا کند ولی بدین شکل نیست زیرا از یک درجه‌ای به بعد مدل کاملا شبیه به داده آموزشی می‌شود و خطا را به صفر می‌رساند و اما داده‌ی تست ما لزومی به شبیه بودن به داده آموزشی ندارد و به همین دلیل خطای واقعی ما ناگهان افزایش چشمگیری می‌یابد و مسئله overfiting در مدل ما اتفاق می‌افتد که سبب این پدیده می‌شود.