Part 9 - Linear Regression

Dalam bab ini, kita akan menerapkan konsep matematika untuk menyelesaikan masalah regresi linear (pencocokan kurva). Tujuan dari regresi adalah untuk menemukan sebuah fungsi f yang memetakan input $x \in \mathbb{R}^D$ ke nilai fungsi yang bersesuaian $f(x) \in \mathbb{R}$. Kita mengasumsikan diberikan satu set data latih yang terdiri dari input x_n dan observasi yang mengandung *noise* (bising)

$$y_n = f(x_n) + \epsilon$$

di mana ϵ adalah variabel acak yang mendeskripsikan *noise* pengukuran. Dalam bab ini, kita mengasumsikan *noise* tersebut berdistribusi Gaussian dengan rata-rata nol. Tugas utamanya adalah menemukan fungsi yang tidak hanya memodelkan data latih dengan baik, tetapi juga dapat melakukan generalisasi untuk memprediksi nilai fungsi pada lokasi input baru yang tidak ada dalam data latih. Beberapa masalah utama yang perlu dipecahkan dalam regresi meliputi pemilihan model, penentuan parameter yang baik, penanganan *overfitting*, pemodelan ketidakpastian, dan memahami hubungan antara fungsi kerugian (*loss*) dan asumsi *prior*.

9.1 Formulasi Masalah

Karena adanya *noise* observasi, kita akan menggunakan pendekatan probabilistik. Secara spesifik, kita memodelkan masalah regresi dengan fungsi *likelihood* (kemungkinan):

$$p(y|x) = \mathcal{N}(y|f(x), \sigma^2)$$

Ini menyiratkan hubungan fungsional $y = f(x) + \epsilon$, di mana $\epsilon \sim \mathcal{N}(0, \sigma^2)$ adalah *noise* Gaussian dengan rata-rata 0 dan varians σ^2. Bab ini berfokus pada model parametrik, di mana parameter $\boldsymbol{\theta}$ muncul secara linear dalam model. Contoh model regresi linear adalah:

$$p(y|x, \boldsymbol{\theta}) = \mathcal{N}(y|x^\top \boldsymbol{\theta}, \sigma^2) \iff y = x^\top \boldsymbol{\theta} + \epsilon$$

Penting untuk dicatat bahwa istilah "regresi linear" merujuk pada model yang linear dalam parameter, bukan harus linear dalam input x. Hal ini memungkinkan kita untuk menerapkan transformasi non-linear $\boldsymbol{\phi}(x)$ pada input dan tetap berada dalam kerangka regresi linear.

9.2 Estimasi Parameter

Diberikan sebuah *training dataset* $\mathcal{D} := (x_1, y_1), \ldots, (x_N, y_N)$, kita bertujuan untuk menemukan parameter optimal $\boldsymbol{\theta}^*$. Karena setiap observasi y_n independen secara kondisional, *likelihood* dari keseluruhan data dapat difaktorkan menjadi produk dari *likelihood* individual:

$$p(\mathbf{Y}|x, \boldsymbol{\theta}) = \prod_{n=1}^{N} p(y_n|x_n, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(y_n|x_n^\top \boldsymbol{\theta}, \sigma^2)$$

9.2.1 Estimasi Kemungkinan Maksimum (*Maximum Likelihood Estimation*)

Maximum Likelihood Estimation (MLE) adalah pendekatan untuk menemukan parameter $\boldsymbol{\theta}_{\text{ML}}$ yang memaksimalkan fungsi *likelihood*. Dalam praktiknya, kita sering kali memaksimalkan logaritma dari *likelihood* (log-*likelihood*) karena transformasi logaritmik tidak mengubah lokasi maksimum dan menyederhanakan perhitungan turunan dengan mengubah produk menjadi penjumlahan. Memaksimalkan log-*likelihood* setara dengan meminimalkan *negative log-likelihood*, yang untuk model regresi linear dengan *noise* Gaussian, dapat ditulis sebagai (mengabaikan konstanta):

$$L(\boldsymbol{\theta}) := \frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - x_n^\top \boldsymbol{\theta})^2 = \frac{1}{2\sigma^2} |\mathbf{y} - x\boldsymbol{\theta}|^2$$

Di sini, x adalah matriks desain di mana setiap baris adalah $x_n^\top$, dan $\mathbf{y}$ adalah vektor yang berisi semua target y_n. Karena fungsi ini kuadratik terhadap $\boldsymbol{\theta}$, solusi global unik dapat ditemukan dengan mengatur gradiennya menjadi nol. Solusi untuk $\boldsymbol{\theta}_{\text{ML}}$ adalah:

$$\boldsymbol{\theta}_{\text{ML}} = (x^\top x)^{-1} x^\top \mathbf{y}$$

Jika kita menggunakan fitur non-linear $\boldsymbol{\phi}(x)$, matriks desain x digantikan oleh matriks fitur $\boldsymbol{\Phi}$, dan solusinya menjadi

$$\boldsymbol{\theta}_{\text{ML}} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{y}$$

Varians *noise* juga dapat diestimasi menggunakan MLE, menghasilkan

$$\sigma^2_{\text{ML}} = \frac{1}{N} \sum_{n=1}^{N} (y_n - \boldsymbol{\phi}(x_n)^\top \boldsymbol{\theta})^2$$

9.2.2 *Overfitting* dalam Regresi Linear

Overfitting terjadi ketika model terlalu fleksibel dan "terlalu cocok" dengan data latih, sehingga gagal menggeneralisasi dengan baik pada data yang tidak terlihat. Contohnya, pada regresi polinomial, penggunaan derajat polinomial yang sangat tinggi ($M \geq N - 1$) akan membuat fungsi melewati setiap titik data latih tetapi berosilasi secara liar di antara titik-titik tersebut. Kualitas model sering dievaluasi menggunakan *Root Mean Square Error* (RMSE) pada data latih dan data uji terpisah. Seiring dengan meningkatnya kompleksitas model (misalnya, derajat polinomial M), *training error* akan terus menurun, tetapi *test error* pada suatu titik akan mulai meningkat, yang menandakan terjadinya *overfitting*.

9.2.3 Estimasi *Maximum A Posteriori* (MAP)

Untuk mengatasi *overfitting*, kita dapat menempatkan distribusi *prior* $p(\boldsymbol{\theta})$ pada parameter, yang mencerminkan keyakinan kita tentang nilai parameter yang masuk akal sebelum melihat data. Alih-alih memaksimalkan *likelihood*, estimasi MAP memaksimalkan distribusi *posterior* $p(\boldsymbol{\theta}|x, \mathbf{Y})$, yang menurut Teorema Bayes sebanding dengan perkalian antara *likelihood* dan *prior*:

$$p(\boldsymbol{\theta}|x, \mathbf{Y}) \propto p(\mathbf{Y}|x, \boldsymbol{\theta})p(\boldsymbol{\theta})$$

Dengan asumsi *prior* Gaussian, $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, b^2\mathbf{I})$, estimasi MAP $\boldsymbol{\theta}_{\text{MAP}}$ ditemukan dengan meminimalkan *negative log-posterior*. Solusinya adalah:

$$\boldsymbol{\theta}_{\text{MAP}} = \left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \frac{\sigma^2}{b^2}\mathbf{I}\right)^{-1}\boldsymbol{\Phi}^\top\mathbf{y}$$

Penambahan suku $\frac{\sigma^2}{b^2}\mathbf{I}$ membantu mencegah *overfitting* dengan "menghukum" nilai parameter yang besar.

9.2.4 Estimasi MAP sebagai Regularisasi

Pendekatan MAP sangat terkait dengan regularisasi. Meminimalkan *negative log-posterior* dengan *prior* Gaussian setara dengan meminimalkan fungsi kerugian *regularized least squares*:

$$|\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}|^2 + \lambda|\boldsymbol{\theta}|_2^2$$

Suku pertama adalah *data-fit term* (berasal dari *likelihood*), dan suku kedua adalah *regularizer* (berasal dari *prior*). Dalam kasus ini, parameter regularisasi λ berhubungan langsung dengan varians dari *likelihood* dan *prior* ($\lambda = \sigma^2/b^2$).

9.3 Regresi Linear Bayesian

Berbeda dengan MLE dan MAP yang menghasilkan estimasi titik tunggal untuk parameter, regresi linear Bayesian menghitung distribusi *posterior* penuh atas parameter, $p(\boldsymbol{\theta}|x, \mathbf{Y})$. Dengan menggunakan *prior* Gaussian $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$ dan *likelihood* Gaussian (model konjugat), distribusi *posterior* yang dihasilkan juga Gaussian:

$$p(\boldsymbol{\theta}|x, \mathbf{Y}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_N, \mathbf{S}_N)$$

dengan *mean* dan kovarians *posterior* diberikan oleh:

$$\begin{aligned}\mathbf{S}_N &= (\mathbf{S}_0^{-1} + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1} \\ \mathbf{m}_N &= \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \sigma^{-2}\boldsymbol{\Phi}^\top\mathbf{y})\end{aligned}$$

Prediksi untuk input baru x_* dibuat dengan melakukan marginalisasi (rata-rata) atas semua kemungkinan parameter sesuai dengan distribusi *posterior* mereka:

$$\begin{aligned}p(y_*|\mathcal{X}, \mathcal{Y}, x_*) &= \int p(y_*|x_*, \boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y})d\boldsymbol{\theta} \\ &= \mathcal{N}(y_*|\boldsymbol{\phi}^\top(x_*)\mathbf{m}_N, \boldsymbol{\phi}^\top(x_*)\mathbf{S}_N\boldsymbol{\phi}(x_*) + \sigma^2)\end{aligned}$$

Hasilnya adalah distribusi prediktif yang juga Gaussian. Varians prediktif ini mencakup dua sumber ketidakpastian: ketidakpastian dari *noise* pengukuran (σ^2) dan ketidakpastian dari parameter ($\boldsymbol{\phi}(x_*)^\top\mathbf{S}_N\boldsymbol{\phi}(x_*)$).

9.4 Kemungkinan Maksimum sebagai Proyeksi Ortogonal

Estimasi *maximum likelihood* memiliki interpretasi geometris yang kuat. Solusi MLE untuk regresi linear, $x\boldsymbol{\theta}_{\text{ML}}$, secara efektif merupakan proyeksi ortogonal dari vektor target observasi $\mathbf{y}$ ke subruang yang direntang oleh kolom-kolom matriks desain x (atau matriks fitur $\boldsymbol{\Phi}$). Dengan kata lain, MLE menemukan vektor dalam subruang model yang memiliki jarak kuadrat terkecil (paling "dekat") dengan data observasi $\mathbf{y}$.