Оцінювання якості моделі машинного навчання, підхід максимуму правдоподібності і байєсівський

Н.І.Недашківська

Інститут прикладного системного аналізу Національного технічного університету України "Київський політехнічний інститут ім. Ігоря Сікорського"

Київ-2021

Основна ідея перевірки якості моделі

Відокремлення перевірочних данних від навчальних.
Розбиття всіх наявних даних на дві (три) підмножини:

- навчальна множина,
- перевірочна / валідаційна множина, на якій перевіряється якість моделі і/або налаштовуються гіперпараметри моделі,
- додатково може бути третя підмножина для остаточного оцінювання якості обраної, "найкращої"моделі.

Один з недоліків використання відокремленого набору даних для перевірки моделі - це *втрата частини даних* для навчання моделі.

Перехресна перевірка моделі МН

Перехресна перевірка (cross-validation) - побудова послідовності класифікацій, в яких кожна підмножина даних використовується як в якості навчальної, так і перевірочної множин. Наприклад, розіб'ємо дані на два набори і по черзі використаємо кожний з них в якості перевірочного (*двохблочна перехресна перевірка*). Далі узагальнити взявши від них *середнє значення*.

У **загальному випадку** дані для перевірки розбиваються на k - кількість блоків, де перевірка виконується на одному блоці, навчання виконується на (k-1)-му блоці.

Помилки навчання, узагальнення і тестування моделі

- Значення функції помилки на навчальному наборі даних називається **помилкою навчання**.
- Здатність алгоритму давати правильний результат на нових даних називається **узагальненням**.
- **Помилка узагальнення/ тестування** - це математичне очікування помилки на нових вхідних даних.
- Припущення: приклади в навчальному і тестовому наборах є незалежними і обидва набори однаково розподілені, вибираються з одного і того ж розподілу ймовірності. Цей загальний розподіл називається **породжуючим розподілом**.

Що означає "якісний"алгоритм МН ?

- помилка навчання якомога менша;
- невеликий розрив між помилками навчання і узагальнення.

Центральні проблеми МН:

- **недостатнє навчання** (недонавчання) - модель не дозволяє отримати досить малу помилку на навчальному наборі,
- **перенавчання** - помилка узагальнення занадто перевищує помилку навчання.

Регуляризація - модифікація алгоритму МН з метою зменшення помилки узагальнення, не зменшуючи помилку навчання.

Недонавчання і перенавчання

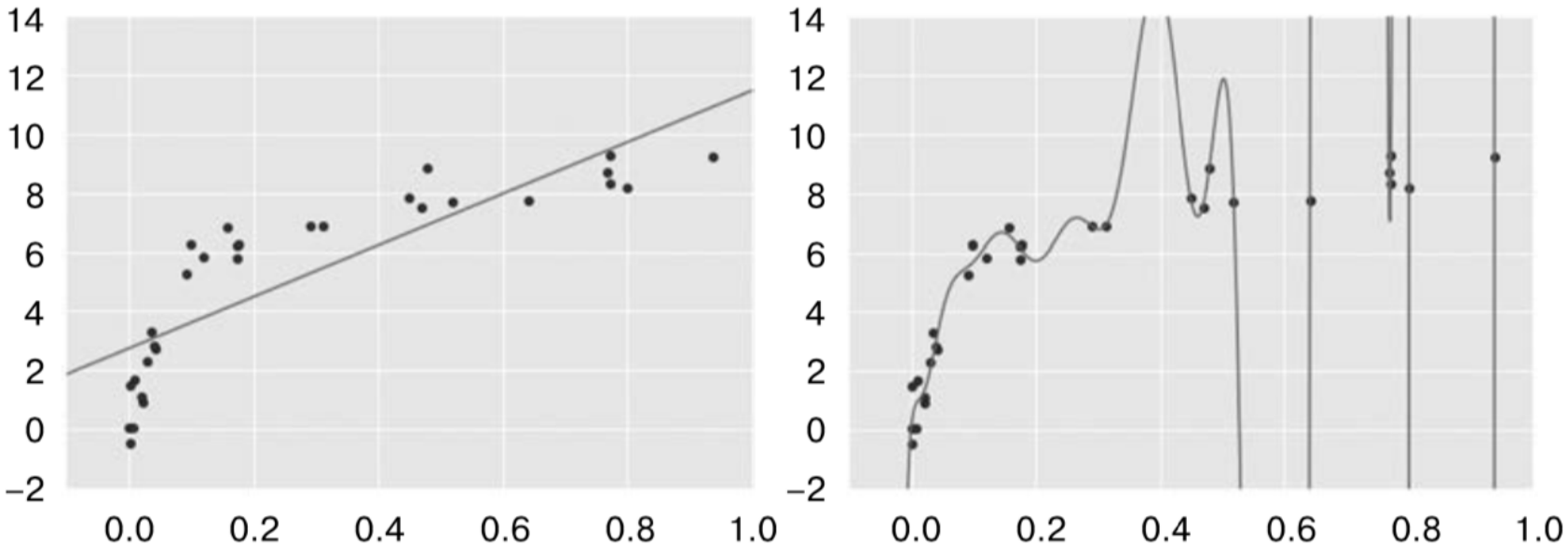


Рис.: Недонавчена модель (лінійна, зліва), перенавчена модель (справа)

Ємність моделі МН

Управляти схильністю моделі до перенавчання або недостатнього навчання дозволяє її **ємність** (capacity).
Простір гіпотез алгоритму - множина функцій, які алгоритм може розглядати в якості потенційного рішення.
Поліноміальна регресія:

$$\hat{y} = b + \sum_{i=1}^{k} w_i x^i,$$

де параметром для управління складністю моделі є ступінь многочлена k.
Яка ємність моделі (ступінь многочлена) забезпечує компроміс між недонавчанням і перенавчанням?

Компроміс між систематичною помилкою і дисперсією моделі

- Для моделей з **великою систематичною помилкою** ефективність моделі на перевірочному наборі даних не набагато гірша за її ефективність на навчальній множині.
- Для моделей з **високою дисперсією** ефективність моделі на перевірочному наборі даних є істотно гіршою за її ефективність на навчальній множині.

Означення систематичної помилки (зміщення)

Нехай навчальні дані $(x_1, x_2, ..., x_m)$ породжуються випадковим процесом.
Точкова оцінка:

$$\hat{w} = h(x_1, x_2, ..., x_m),$$

де x_i - незалежні однаково розподілені точки.
$\hat{w}$ - випадкова величина.
Систематична помилка (зміщення, bias):

$$bias(\hat{w}) = M(\hat{w}) - w,$$

$M(\hat{w})$ - вибіркове математичне сподівання, w - істинне значення параметра.
Оцінка називається незміщеною, якщо

$$M(\hat{w}) = w$$

.

Дисперсія середнього значення помилки

- Бажані оцінки - які мають малі зміщення і дисперсію.
- Оцінка середнього значення помилки на основі вибірки скінченого розміру - недостовірна.
- Дисперсія для середнього значення:

$$Var(\hat{\mu}) = Var\left(\frac{1}{m}\sum_i x_i\right) = \frac{\sigma^2}{m},$$

σ^2 - істинна дисперсія вибірки x_i, яка має нормальний розподіл.

- Стандартна помилка для середнього значення: $SE(\hat{\mu}) = \frac{\sigma}{\sqrt{m}}$ часто оцінюється за допомогою оцінки σ - кореня з *незміщеної* оцінки дисперсії. При великих m оцінка вважається прийнятною.

Дисперсія середнього значення помилки (продовження)

- **Оцінка помилки узагальнення** моделі МН - вибіркове середнє помилки на тестовій множині.
- Розмір тестової множини даних і визначає точність цієї оцінки.
- Згідно з центральною граничною теоремою середнє значення має наближено нормальний розподіл.
- Тому стандартну помилку $SE(\hat{\mu})$ використовують для обчислення імовірності, що істинне математичне сподівання помилки знаходиться у вибраному інтервалі, наприклад, у 95% довірчому інтервалі:

$$(\hat{\mu} - 1.96SE(\hat{\mu}), \hat{\mu} + 1.96SE(\hat{\mu}))$$

при нормальному розподілі з параметрами $\hat{\mu}$ і $SE(\hat{\mu})^2$.

Дисперсія середнього значення помилки (продовження)

Стандартну помилку середнього $SE(\hat{\mu}) = \frac{\sigma}{\sqrt{m}}$ можна використати для обчислення імовірності, що істинне математичне сподівання помилки знаходиться у вибраному інтервалі, наприклад, у 95% довірчому інтервалі:

$$(\hat{\mu} - 1.96SE(\hat{\mu}), \hat{\mu} + 1.96SE(\hat{\mu}))$$

при нормальному розподілі з середнім $\hat{\mu}$ і дисперсією $SE(\hat{\mu})^2$. **Алгоритм МН А вважається кращим за алгоритм В**, якщо довірчий інтервал для середнього значення помилки алгоритму А строго менший за довірчий інтервал для середнього значення помилки алгоритму В.

Компроміс між систематичною помилкою і дисперсією

Вибір «оптимальної моделі» полягає у відшуканні найкращого компромісу між систематичною помилкою (зміщенням, bias) та дисперсією.

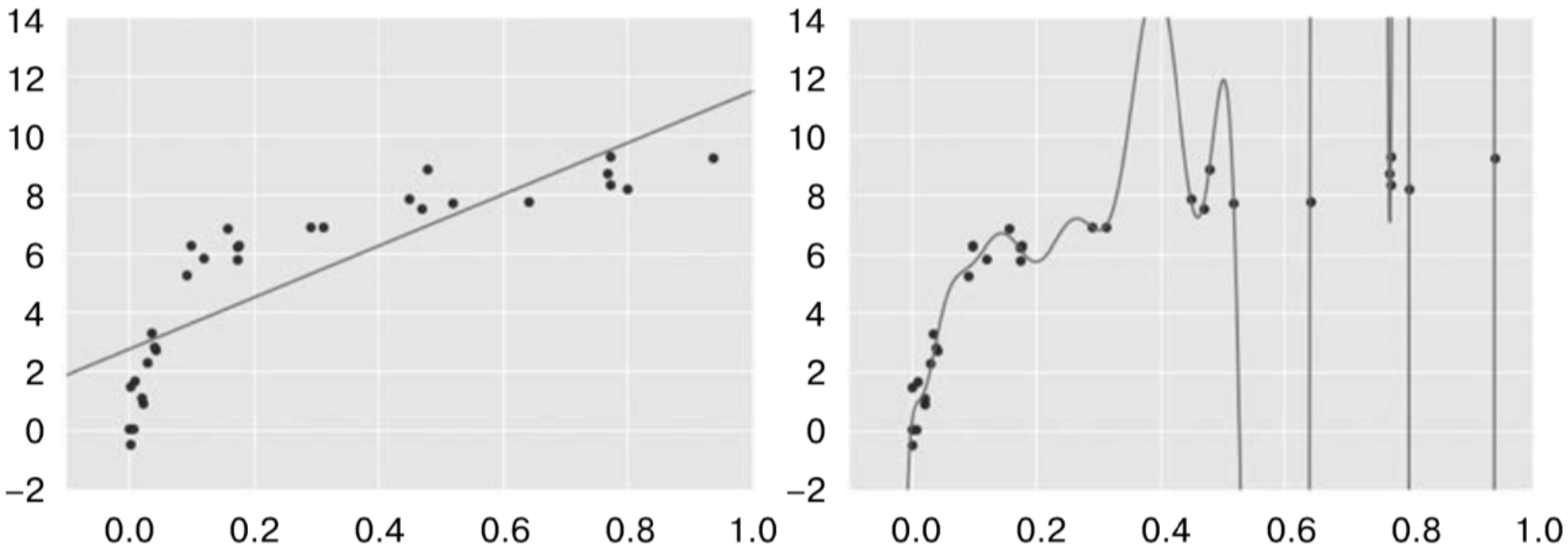


Рис.: Моделі регресії з великою систематичною помилкою (лінійна, зліва) і великою дисперсією (справа)

Крива перевірки в Scikit-Learn

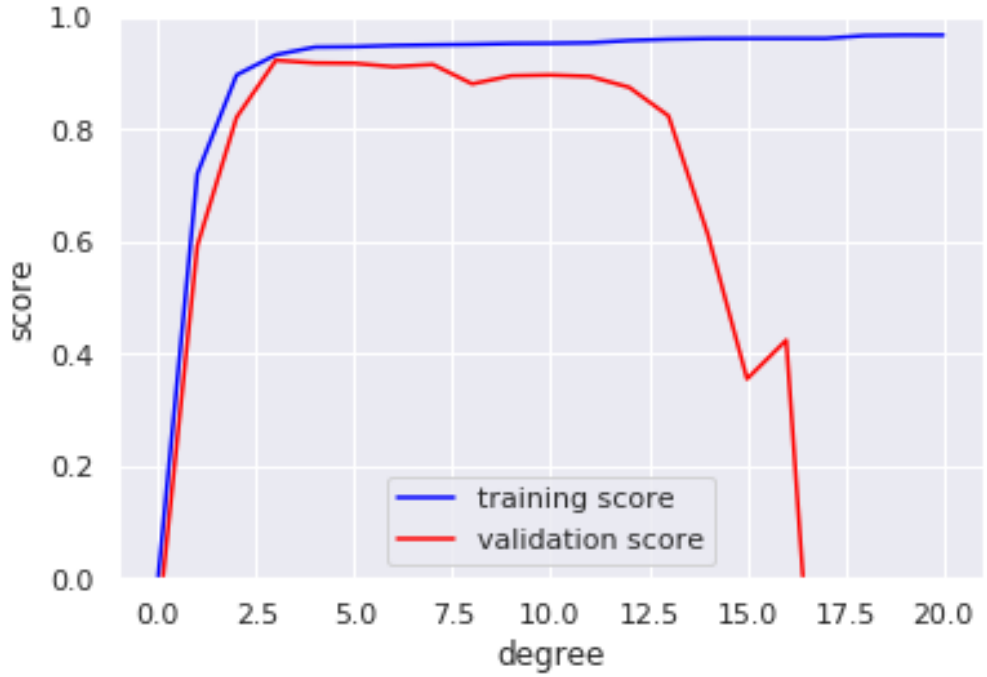


Рис.: Крива перевірки (validation score), degree - ступінь многочлена в моделі поліноміальної регресії, training score - оцінка навчання

Для даного прикладу компроміс між систематичною помилкою і дисперсією досягається для многочлена третього ступеня.

Висновки на основі кривої перевірки

- Ефективність моделі на навчальній множині зазвичай вища за ефективність на перевірочній множині.
- Оцінка ефективності на навчальній множині монотонно зростає із ростом складності моделі.
- Моделі з низькою складністю недостатньо навчені, погано прогнозують як дані навчальної множини, так і нові дані.
- Моделі з високою складністю (високою дисперсією) перенавчені, добре прогнозують дані навчальної множини, на нових даних працюють погано.
- Крива перевірки досягає максимуму в деякій проміжній точці: цей рівень складності означає прийнятний компроміс між систематичною помилкою і дисперсією.

Компроміс між систематичною помилкою і дисперсією

Зміщення і дисперсія – це два джерела помилки оцінки. Нехай маємо дві оцінки, у однієї велике зміщення, а у іншої велика дисперсія.

Так яку оцінку вибрати?

Підходи до пошуку компромісу:

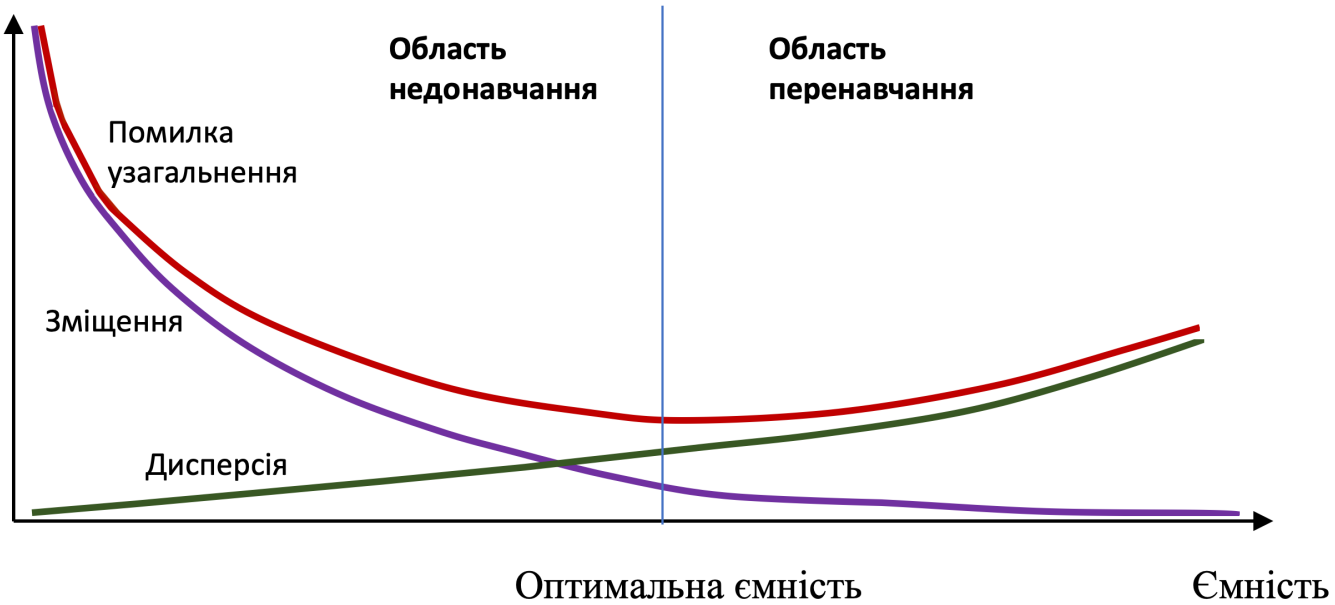
1. Перехресна перевірка.
2. Порівняти середньоквадратичну помилку MSE обох оцінок (для задачі регресії):

$$MSE = M((\hat{w} - w)^2) = bias(\hat{w}) + Var(\hat{w}).$$

Бажаною є оцінка з малою MSE.

Компроміс між систематичною помилкою і дисперсією

U-подібна крива залежності помилки узагальнення від ємності моделі:



Як себе веде оцінка при зростанні розміру навчальної вибірки?

Конзистентність оцінки:
$\hat{w} \to w$, $\forall w \in W$ **за імовірністю** при $m \to \infty$,
m - розмірність навчальної вибірки:

$$\forall \epsilon > 0 : P(|\hat{w} - w| > \epsilon) \to 0, m \to \infty.$$

1. Для конзистентної оцінки: зміщення оцінки зменшується з ростом числа прикладів.
   Оцінка називається асимптотично незміщенною, якщо
   $$\lim_{m\to\infty} bias(\hat{w}_m) = 0, \lim_{m\to\infty} M(\hat{w}_m) = w.$$
2. З асимптотичної незміщеності НЕ випливає конзистентність.

Оцінка максимальної правдоподібності

Розглянемо навчальну множину прикладів $X = (x_1, x_2, ..., x_m)$, які незалежно вибираються з невідомого породжуючого розподілу $p_{data}(x)$.
Позначимо $p_{model}(x; \theta)$ - параметричне сімейство розподілів ймовірності над одним і тим же простором, ці розподіли визначаються параметром θ.
Оцінка максимальної правдоподібності для θ:

$$\theta_{ML} = \arg\max_{\theta} p_{model}(X; \theta) =$$

$$= \arg\max_{\theta} \prod_{i=1}^{m} p_{model}(x_i; \theta).$$

Оцінка максимальної правдоподібності (продовження)

Для отримання більш зручної задачі беруть логирифм від оптимізаційної функції:

$$\theta_{ML} = \arg\max_{\theta} \sum_{i=1}^{m} \ln p_{model}(x_i; \theta).$$

Розділимо праву частину на m:

$$\theta_{ML} = \arg\max_{\theta} M_{x \sim \hat{p}_{data}} \ln p_{model}(x; \theta).$$

Оцінка максимальної правдоподібності (продовження)

Максимальна правдоподібність - це спроба наблизити модельний розподіл до емпіричного розподілу $\hat{p}_{data}$.
Розходження Кульбака-Лейблера - метрика між двома розподілами:

$$D_{KL}(\hat{p}_{data}||p_{model}) = M_{x\sim\hat{p}_{data}}(\ln \hat{p}_{data}(x) - \ln p_{model}(x)).$$

$\hat{p}_{data}$ - емпіричний розподіл, який визначається навчальною вибіркою,
p_{model} - модельний розподіл.
Мінімізація розходження КЛ зводиться до:

$$-M_{x\sim\hat{p}_{data}}(\ln p_{model}(x)) \to \min.$$

Властивості оцінки максимальної правдоподібності

Вона є **асимптотично найкращою оцінкою** з точки зору швидкості збіжності, коли $m \to \infty$.

1. Вона є конзистентною за умов:
   - істинний розподіл p_{data} належить сімейству модельних розподілів $p_{model}(;\theta)$,
   - істинний розподіл p_{data} відповідає рівно одному значенню θ.
2. Вона є статистично ефективною - дає меншу помилку узагальнення при фіксованому числі прикладів m.

Властивості оцінки максимальної правдоподібності (продовження)

Статистична ефективність вивчається в задачах оцінювання значення параметра (напр, в задачі регресії). Для великих m справедлива нерівність Крамера-Рао, яка показує, що ні для якої конзистентної оцінки середньоквадратична помилка не може бути меншою ніж для оцінки максимальної правдоподібності.
Оцінці максимальної правдоподібності часто віддають перевагу в машинному навчанні.

Умовна логарифічна правдоподібність

Будемо оцінювати умовну імовірність $p(y|x; w)$.
Нехай X - всі входи, Y - всі спостережувані виходи. Оцінка умовної логарифічної правдоподібності:

$$w_{ML} = \arg\max_w p(Y|X; w).$$

Якщо всі елементи навчальної вибірки однаково розподілені та незалежні, то:

$$w_{ML} = \arg\max_w \sum_{i=1}^{m} \ln p(y_i|x_i; w).$$

Максимальна правдоподібність та лінійна регресія

Чому саме для моделі $\hat{y} = w^T x$ використовується

$$MSE(w) = \frac{1}{m}\sum_{j=1}^{m}(y_j - \hat{y}_j)^2?$$

Будуємо модель, яка буде породжувати умовний розподіл $p(y|x)$.
Модель лінійної регресії базується на припущенні:

$$z = \hat{y}(x, w) + \epsilon, \epsilon \sim N(0, \sigma^2),$$

σ^2 - фіксована константа.
Тоді спостережувана змінна також нормально розподілена:

$$p(z|x, w, \sigma^2) = N(z|\hat{y}(x, w), \sigma^2).$$

Є навчальна вибірка $X = \{(x_j, z_j)\}, j = 1, ..., m$. Нехай z_j незалежні, однаково розподілені випадкові величини:

Максимальна правдоподібність та лінійна регресія

Будемо максимізувати логарифічну правдоподібність відносно w.

$$\ln p(z|X, w, \sigma^2) = \sum_{j=1}^{m} \ln p(z_j|\hat{y}(x_j, w), \sigma^2) =$$

$$= \sum_{j=1}^{m} \left(\ln(\frac{1}{\sigma\sqrt{2\pi}}) + \ln(\exp(-\frac{(z_j - \hat{y}(x_j, w))^2}{2\sigma^2})) \right).$$

Після перетворень отримаємо (перевірити !!!),

$$\ln p(w|X) = \left(-\frac{m}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{j=1}^{m} (z_j - x_j^T w)^2 \right),$$

Максимізація $\ln p(w|X)$ потребує мінімізації середньоквадратичної помилки, що і треба було показати.

Байесівський підхід до оцінювання параметрів

- Прогноз робиться за результатами розгляду **всіх можливих значень** параметрів w.
- Імовірність відображає **ступінь впевненості** в наших знаннях.
- До спостереження даних, знання представляються деяким **апріорним** розподілом імовірності з великою ентропією (напр., рівномірним або нормальним розподілом).
- Використовується **теорема (правило) Байеса**.
- Після спостереження даних, розраховується **апостеріорний розподіл** з меншою ентропією, сконцентрований в околі імовірних значень параметрів.

Теорема Байеса

Нехай $X = (x_1, x_2, ..., x_m)$ - навчальна множина прикладів.
Теорема Байеса:

$$p(w|X) = \frac{p(w)p(X|w)}{p(X)},$$

- $p(w)$ - щільність апріорної імовірності,
- $p(X|w)$ - правдоподібність,
- $p(w|X)$ - щільність апостеріорної імовірності,
- $p(X) = \int_{w \in W} p(w)p(X|w)dw$ - імовірність даних.

Відмінності між оцінкою максимальної правдоподібності (МП) та байесівською (Б)

- Прогноз в підході МП виконується на основі точкової оцінки параметра, в Б. підході - на основі повного розподілу.
  Наприклад, розподіл наступного прикладу x_{m+1} після спостереження m прикладів:

$$p(x_{m+1}|x_1, x_2, ..., x_m) = \int p(x_{m+1}|w)p(w|x_1, x_2, ..., x_m)dw.$$

  Кожне значення параметру w вносить вклад в прогнозування наступного прикладу.
- Вклад апріорного розподілу в Б.підході.

Максимальна апостеріорна гіпотеза

Максимальна апостеріорна гіпотеза (maximum a posteriori hypothesis, MAP):

$$w_{MAP} = \arg\max_w p(w|X) = \arg\max_w p(w)p(X|w).$$

- $p(w)$ - щільність апріорної імовірності,
- $p(X|w)$ - правдоподібність,

$$w_{MAP} = \arg\max_w \ln p(w|X) = \arg\max_w \ln p(w) + \ln p(X|w).$$

- $\ln p(X|w)$ - стандартна логарифмічна правдоподібність

Байесівський підхід до регуляризації зі зниженням ваги в задачі регресії

Вводиться апріорний розподіл для вектора ваг:

$$p(w) = N(w|\mu_0, \sigma_0^2).$$

Нехай в наборі даних $X = \{(x_j, z_j), j = 1, ..., m\}$ величини z_j незалежні випадкові, однаково розподілені за нормальним законом:

$$p(z|X, w, \sigma^2) = \prod_{j=1}^{m} N(z_j|x_j^T w, \sigma^2).$$

Байесівський підхід до регуляризації зі зниженням ваги

Використаємо формулу Байеса, щоб знайти апостеріорний розподіл для вектора ваг:

$$p(w|z) = \frac{p(w)p(z|w)}{p(z)} = \frac{1}{p(z)} N(w|\mu_0, \sigma_0^2) \prod_{j=1}^{m} N(z_j|x_j^T w, \sigma^2)$$

Нехай апріорний розподіл для вектора ваг:

$$p(w) = N(w|0, \frac{1}{\lambda} I),$$

де 0 - нульовий вектор, $\frac{1}{\lambda} I$ - коваріаційна матриця, I - одинична матриця.

Байесівський підхід до регуляризації зі зниженням ваги

Тоді логарифм апостеріорної імовірності (ВПРАВА !!!):

$$\ln p(w|z) = -\frac{1}{2\sigma^2}\sum_{j=1}^{m}(z_j - x_j^T w)^2 - \frac{\lambda}{2}w^T w + c.$$

- Максимізація $\ln p(w|z)$ потребує мінімізації функції

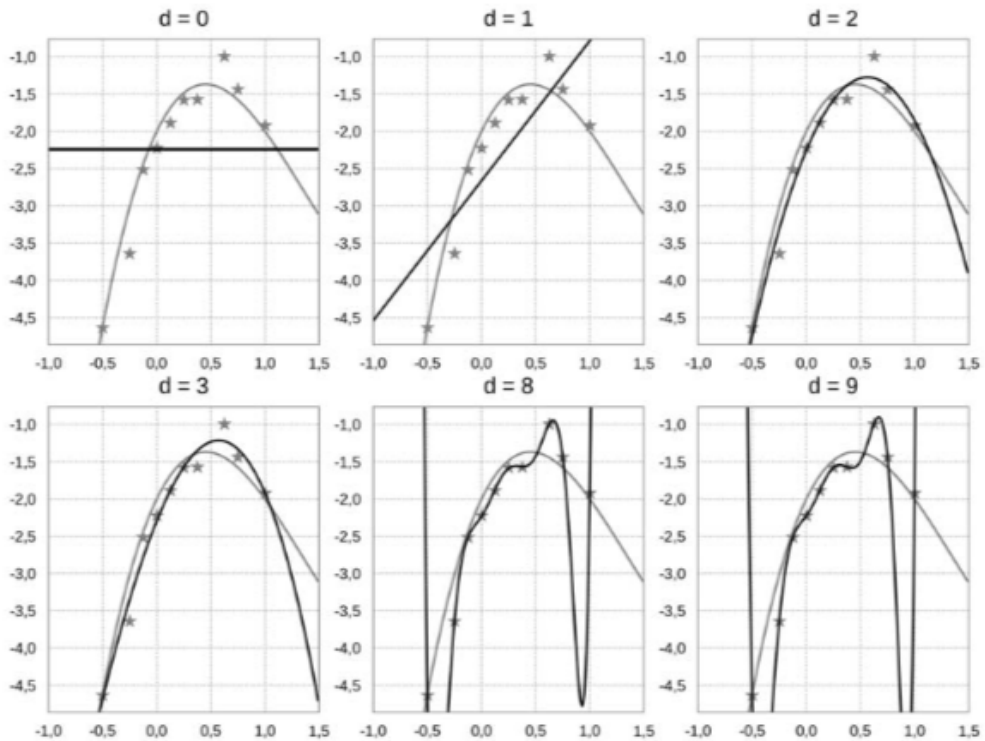
$$L(w) = \frac{1}{2}\sum_{j=1}^{m}(z_j - x_j^T w)^2 + \frac{\lambda}{2}||w||^2,$$

.

- За умови нормального апріорного розподілу $p(w)$ і нормального розподілу $p(z|X, w, \sigma^2)$ з постійною дисперсією, логарифм апостеріорного розподілу $p(w|z)$ - це квадратична функція від w, тобто $p(w|z)$ також має нормальний розподіл.

Регуляризація зі зниженням ваги. Приклад

Точки - значення многочлена $f(x) = x^3 - 4x^2 + 3x - 2$, до яких додано шум $\epsilon \sim N(0, 1/4)$.



Регуляризація зі зниженням ваги. Приклад

Приклад: будемо шукати многочлен степені k

$$y(x) = \sum_{j=0}^{k} x^j w_j,$$

який оптимальним чином описує дані з навчальної вибірки $\{(x_j, y_j)\}, j = 1, ..., m$. Випишемо рівняння многочленів k-го ступеня, за допомогою яких виконувалося наближення даних в прикладі: $f_0(x) = -2.2393$,
$f_1(x) = -2.6617 + 1.8775x$,
$f_2(x) = -2.2528 + 3.4604x - 3.0603x^2$,
$f_8(x) = -2.23 + 2.23x + 6.25x^2 + 15.60x^3 - 239.98x^4 +$
$+322.85x^5 + 621.09x^6 - 1478.65x^7 + 750.90x^8$,
$f_9(x) = -2.22 + 2.01x + 4.88x^2 + 31.13x^3 - 230.31x^4 +$
$+103.72x^5 + 869.22x^6 - 966.67x^7 - 319.31x^8 + 505.64x^9$.

Регуляризація зі зниженням ваги

Хочемо обмежити норму вектора ваг w.
Додамо у функцію середньоквадратичної помилки додатковий доданок (**регуляризатор**)

$$L(w) = \frac{1}{2}\sum_{j=1}^{m}(y_j - x_j^T w)^2 + \frac{\lambda}{2}||w||^2,$$

який відповідає за розмір коефіцієнтів у многочленів.
Параметр λ називається **коефіцієнтом регуляризації**.
Метод регуляризації, в якому до функції помилки додається доданок $\frac{\lambda}{2}||w||^2$, називається **гребневою регресією (ridge regression)**.

Регуляризація зі зниженням ваги. Гребнева регресія

Виконаємо перетворення в функції $L(w)$, отримаємо

$$L(w) = \frac{1}{2}(y - Xw)^T(y - Xw) + \frac{\lambda}{2}w^T w.$$

Виконаємо диференціювання $L(w)$ за w, отримаємо (ВПРАВА !!!):

$$w^* = (X^T X + \lambda I)^{-1} X^T y,$$

де I - одинична матриця.

Регуляризація зі зниженням ваги. Приклад

Додамо регуляризатор в нашому прикладі до многочлена дев'ятого степеня. Отримаємо наступні многочлени:

$$f_{\lambda=0}(x) = -2.22 + 2.01x + 4.88x^2 + 31.13x^3 - 230.31x^4 +$$

$$+103.72x^5 + 869.22x^6 - 966.67x^7 - 319.31x^8 + 505.64x^9.$$

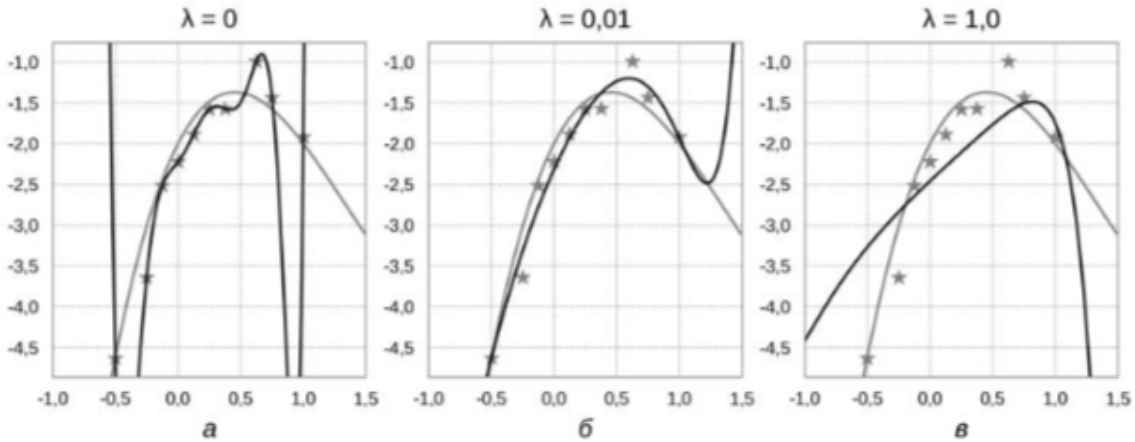
$$f_{\lambda=0.01}(x) = -2.32 + 3.40x - 2.33x^2 + 0.05x^3 - 0.51x^4 -$$

$$-0.29x^5 - 0.22x^6 - 0.06x^7 + 0.09x^8 + 0.24x^9.$$

$$f_{\lambda=1}(x) = -2.46 + 1.45x - 0.19x^2 + 0.22x^3 - 0.13x^4 -$$

$$-0.05x^5 - 0.14x^6 - 0.13x^7 - 0.16x^8 - 0.16x^9.$$

Регуляризація зі зниженням ваги. Приклад



Прийнятні результати маємо при $\lambda = 0.01$

Дякую за увагу!