natychmiast widzimy, że maksymalizacja logarytmicznej wiarygodności względem w daje w wyniku tę samą estymację parametrów $\boldsymbol{w}$, jak to robi minimalizacja błędu średniokwadratowego. Oba kryteria mają różne wartości, ale to samo położenie optimum. Uzasadnia to korzystanie z MSE jako procedury estymacji maksymalnej wiarygodności. Jak zobaczymy, estymator maksymalnej wiarygodności ma kilka pożądanych właściwości.

5.5.2. Właściwości maksymalnej wiarygodności

Podstawową zaletą estymatora maksymalnej wiarygodności jest fakt, że można go pokazać jako najlepszy estymator asymptotycznie, jako liczbę przykładów $m \to \infty$, w sensie jego stopnia zbieżności w miarę wzrostu m.

Przy odpowiednich warunkach estymator maksymalnej wiarygodności ma właściwość spójności (patrz punkt 5.4.5), co oznacza, że w miarę jak liczba przykładów szkoleniowych zbliża się do nieskończoności, estymacja maksymalnej wiarygodności jest zbieżna do prawdziwej wartości parametru. Te warunki są następujące:

- prawdziwy rozkład p_{data} musi leżeć w obrębie rodziny modeli $p_{model}(\cdot; \boldsymbol{\theta})$; w przeciwnym przypadku żaden estymator nie może odtworzyć p_{data};
- prawdziwy rozkład p_{data} musi odpowiadać dokładnie jednej wartości $\boldsymbol{\theta}$; w przeciwnym przypadku maksymalna wiarygodność pozwala odtworzyć poprawne p_{data}, ale nie będzie w stanie określić, która z wartości $\boldsymbol{\theta}$ była używana w procesie generowania danych.

Można stąd wyprowadzić inne zasady, poza estymatorem maksymalnej wiarygodności. Wiele z nich ma tę samą właściwość: są estymatorami spójnymi. Spójne estymatory mogą jednak różnić się od siebie **efektywnością statystyczną**, co oznacza, że mogą dawać niższy błąd uogólnienia dla stałej liczby próbek m, albo ekwiwalentnie mogą wymagać mniej przykładów do uzyskania stałego poziomu błędu uogólnienia.

Efektywność statystyczna jest zwykle analizowana dla **przypadku parametrycznego** (jak w regresji liniowej), gdzie naszym celem jest estymacja wartości parametru (przy założeniu, że można zidentyfikować prawdziwy parametr), a nie wartości funkcji. Sposobem zmierzenia, jak blisko prawdziwych parametrów się znajdujemy, jest oczekiwany błąd średniokwadratowy, obliczany jako podniesiona do kwadratu różnica między wartościami estymowanej i prawdziwej wartości parametru, gdzie oczekiwanie dotyczy m próbek

szkoleniowych z rozkładu generującego dane. Ten parametryczny błąd średniokwadratowy maleje ze wzrostem m, a dla dużych m kres dolny Rao-Craméra (Rao 1945, Cramér 1946) pokazuje, że żaden spójny estymator nie ma niższej wartości MSE niż estymator maksymalnej wiarygodności.

Z tych powodów (spójności i efektywności) maksymalna wiarygodność jest często traktowana jako preferowany estymator w systemach uczących się. Gdy liczba przykładów jest na tyle mała, że powoduje zachowania związane z nadmiernym dopasowaniem, strategie regularyzacyjne, jak zanikanie wagi, mogą być używane do uzyskania obciążonej wersji maksymalnej wiarygodności, która ma mniejszą wariancję przy ograniczonych danych szkoleniowych.

5.6. Statystyki Bayesa

Dotąd omawialiśmy **statystykę częstościową** oraz podejście oparte na estymacji pojedynczej wartości $\boldsymbol{\theta}$, a więc wszelkie prognozy były oparte na jednej estymacji. Innym podejściem jest rozważenie przy tworzeniu prognozy wszystkich możliwych wartości $\boldsymbol{\theta}$. Jest ono dziedziną **statystyki Bayesa**.

Jak to omówiono w punkcie 5.4.1, perspektywa częstościowa polega na tym, że prawdziwa wartość parametru $\boldsymbol{\theta}$ jest ustalona, lecz nieznana, natomiast estymacja punktu $\hat{\boldsymbol{\theta}}$ jest zmienną losową, z uwagi na to, że jest funkcją zbioru danych (który jest uważany za losowy).

Bayesowskie spojrzenie na statystykę jest całkiem inne. Wykorzystuje ono prawdopodobieństwo jako odzwierciedlenie stopnia pewności stanu wiedzy. Zbiór danych jest bezpośrednio obserwowany, więc nie jest losowy. Z drugiej strony prawdziwa wartość parametru $\boldsymbol{\theta}$ jest nieznana lub niepewna, a więc jest reprezentowana jako zmienna losowa.

Zanim zaobserwujemy dane, zaprezentujmy naszą wiedzę na temat $\boldsymbol{\theta}$, wykorzystując **rozkład aprioryczny** $p(\boldsymbol{\theta})$ (czasami określany po prostu jako aprioryczny). Ogólnie praktycy systemów uczących się wybierają taki rozkład wstępny, który jest dość szeroki (tj. o wysokiej entropii), aby odzwierciedlić wysoki stopień niepewności co do wartości $\boldsymbol{\theta}$, zanim zaobserwują jakiekolwiek dane. Na przykład można założyć *a priori*, że $\boldsymbol{\theta}$ leży w pewnym skończonym zakresie lub wielkości, o rozkładzie jednostajnym. Wiele rozkładów wstępnych odzwierciedla natomiast preferencje dla „prostszych" rozwiązań (jak współczynniki mniejszej wielkości lub funkcja bliższa stałej).

Przyjmijmy teraz, że mamy zbiór próbek danych $\{x^{(1)}, \ldots, x^{(m)}\}$. Możemy ustalić wpływ danych na nasze przekonanie na temat $\boldsymbol{\theta}$, łącząc ze sobą wiarygodność danych $p(x^{(1)}, \ldots, x^{(m)} \mid \boldsymbol{\theta})$ z prawdopodobieństwem wstępnym