Selected Topics for Derivation

Robert Schmidt

Abstract

The following are notes on the key results from the **Elements of Statistical Learning** text. They were primarily derived from course notes and readings in the Stanford STATS 315: *Modern Applied Statistics* series.

Contents

1 Splines

1.1 Derivation: piecewise polynomials and splines

1.2 Interpolating natural spline minimizes smoothing spline problem

Suppose $N \geq 2$, with $g(x)$ as a natural cubic spline that interpolates $\{x_i, z_i\}_{i=1}^{N}$, with $a < x_1 < \cdots < x_N < b$. Let $\tilde{g}$ be any other differentiable function on $[a, b]$ that interpolates the N pairs. For this derivation, I follow the outline set forth in ESL problem 5.7.

1.2.1 Step 1: Integral result

Let $h(x) = \tilde{g}(x) - g(x)$. Integrating by parts, we see that

$$\int_a^b g''(x)h''(x)\,dx = g''(x)h'(x)\Big|_a^b - \int_a^b g'''(x)h'(x)\,dx$$

Since $g(x)$ is a natural spline: $g''(a) = 0 = g''(b) \implies g''(x)h'(x)\big|_a^b = 0$.
So, we see that $\int_a^b g''(x)h''(x)\,dx = -\int_a^b g'''(x)h'(x)\,dx$.
We can break $-\int_a^b g'''(x)h'(x)\,dx$ into knots given that the spline is defined piecewise:

$$-\int_a^b g'''(x)h'(x)\,dx = -\sum_{j=1}^{N-1} \int_{x_j}^{x_{j+1}} g'''(x)h'(x)\,dx$$

Furthermore, we can do each of the integrals in the sum separately, and integrate each one by parts:

$$-\sum_{j=1}^{N-1} \int_{x_j}^{x_{j+1}} g'''(x)h'(x)\,dx = -\sum_{j=1}^{N-1} g'''(x)h(x)\Big|_{x_j}^{x_{j+1}} + \sum_{j=1}^{N-1} \int_{x_j}^{x_{j+1}} g^{(4)}(x)h(x)\,dx$$

Since $g(x)$ is piecewise cubic, $g^{(4)}(x) = 0\ \forall\, x$.
Summarizing all steps up to this point, $\int_a^b g''(x)h''(x)\,dx = -\sum_{j=1}^{N-1} g'''(x)h(x)\big|_{x_j}^{x_{j+1}}$.
Again noting that $g(x)$ is piecewise cubic, we can rewrite the right hand side expression:

$$-\sum_{j=1}^{N-1} g'''(x)h(x)\Big|_{x_j}^{x_{j+1}} = -\sum_{j=1}^{N-1} g'''(x_j^+)(h(x_{j+1}) - h(x_j))$$

We now consider $\int_a^b g''(x)h''(x)\,dx = -\sum_{j=1}^{N-1} g'''(x_j^+)(h(x_{j+1}) - h(x_j))$.
Recalling the definition of $h(x) = \tilde{g}(x) - g(x)$, we see that $\tilde{g}(x_i) = g(x_i)$ at each endpoint x_i (they are both interpolating functions) $\implies h(x_i) = 0$ for all endpoints.
So, $\sum_{j=1}^{N-1} g'''(x_j^+)(h(x_{j+1}) - h(x_j)) = 0$.

$$\boxed{\therefore \int_a^b g''(x)h''(x)\,dx = 0}$$

1.2.2 Step 2: Inequality result

In step 1, we showed that $\int_a^b g''(x)h''(x)\,dx = 0$.
We now consider $\int_a^b \tilde{g}''(t)^2\,dt$.

$$\begin{aligned}
\int_a^b \tilde{g}''(t)^2\,dt &= \int_a^b (h''(t) + g''(t))^2\,dt \quad \text{using the definition of } h(x) \\
&= \int_a^b h''(t)^2\,dt + \int_a^b g''(t)^2\,dt + 2\int_a^b h''(t)g''(t)\,dt \\
&= \int_a^b h''(t)^2\,dt + \int_a^b g''(t)^2\,dt + 0 \quad \text{by 5.7 (a)} \\
&= \int_a^b h''(t)^2\,dt + \int_a^b g''(t)^2\,dt
\end{aligned}$$

This would imply that $\boxed{\int_a^b \tilde{g}''(t)^2\,dt \geq \int_a^b g''(t)^2\,dt}$ $\forall\, t \in [a,b]$ since $h''(t)^2 \geq 0$ everywhere.
Note that $h''(t) = 0$ everywhere would imply that $h(x)$ is linear on $[a,b]$.
Given that $h(x_i) = 0$ for all knots/endpoints x_i and $N \geq 2$, this could only be true if $h = 0\ \forall\, t \in [a,b]$, which would then imply that $g(x) = \tilde{g}(x)$.
So, equality only holds if h is identically zero in $[a,b]$.

1.2.3 Step 3: Conclusion on minimizer

We now consider the penalized least squares problem:

$$\min_f \left[\sum_{i=1}^N \big(y_i - f(x_i)\big)^2 + \lambda \int_a^b f''(t)^2\,dt \right]$$

For a minimizer $f = \tilde{g}$, we can construct a natural cubic spline g with the same values as $\tilde{g}$ at the spline's knots, $\{x_i\}_{i=1}^N$.
This implies that $\sum_{i=1}^N (y_i - \tilde{g}(x_i)) = \sum_{i=1}^N (y_i - g(x_i))$.
Since $\tilde{g}$ is a minimizer, and due to the step 2 result (in the case that $h(x) = 0$ everywhere), we know that

$$\lambda \int_a^b \tilde{g}''(t)^2\,dx = \lambda \int_a^b g''(t)^2\,dx$$

This implies that $f = \tilde{g} = g$, and therefore that $\boxed{\text{the natural cubic spline is the minimizer.}}$

2 Smoothing matrices

2.1 Reinsch form and kernel matrix

2.2 Proof of kernel trick

3 Semi-parametric linear modeling

3.1 Formulation

3.2 Solution

3.3 Solution properties

4 Trees

4.1 Tree estimate

4.2 Improvement in loss by splitting

5 Neural network backprop overview