CS-567 Machine Learning

Nazar Khan

PUCIT

Lecture 08
Optimization

Model Selection

- In our polynomial fitting example, $M = 3$ gave the best generalization by controlling the number of free parameters.
- Regularization coefficient λ also achieves a similar effect.
- Parameters such as λ are called **hyperparameters**.
- They determine the model (model's complexity).
- Model selection involves finding the best values for parameters such as M and λ.

Model Selection

- One approach is to check generalization on a separate **validation set**.
- Select model that performs best on validation set.
- One standard technique is called **cross-validation**.
  - Use $\frac{S-1}{S}$ of the available data for training and the rest for validation.
  - Disadvantage: S times more training for 1 parameter. S^k times more training for k parameters.

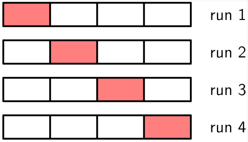


Figure: S-fold cross validation for $S = 4$. Every training is evaluated on the validation set (in red) and these validation set performance are averaged over the S training runs.

Model Selection

- Ideally
  - use only training data,
  - perform only 1 training run for multiple hyperparameters,
  - performance measure that avoids bias due to over-fitting.

Model Selection

- Choose model for which

$$\ln p(\mathcal{D}|\mathbf{w}_{ML}) - M$$

  is maximized.
- This is called **Akaike Information Criterion (AIC)**.
- **The best method is the Bayesian approach which penalises model complexity in a natural, principled way.**

Curse of Dimensionality

- Our polynomial curve fitting example was for a single variable x.
- When number of variables increases, the number of parameters increases exponentially.

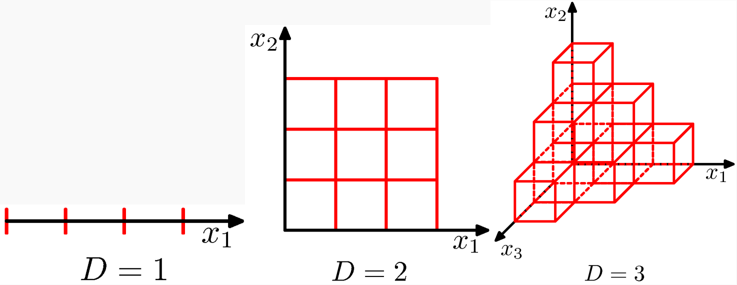


Figure: Curse of Dimensionality: The number of regions of a regular grid grows exponentially with with the dimensionality D of the search space.

Calculus of Variations

Calculus of Real Numbers

- Considers real-valued functions $f(x)$: mappings from a real number x to another real number.
- If f has a minimum in ξ, then ξ necessarily satisfies $f'(\xi) = 0$.
- If f is strictly convex, then ξ is the unique minimum.

Calculus of Variations
Calculus of Variations

- Considers real-valued **functionals** $E(u)$: mappings from a function $u(x)$ to a real number
- If E is minimised by a function v, then v necessarily satisfies the corresponding **Euler-Lagrange** equation, a differential equation in v.
- If E is strictly convex, then v is the unique minimiser.

Calculus of Variations

Euler-Lagrange Equation in 1-D

A smooth function $u(x), x \in [a, b]$ that minimises the functional

$$E(u) = \int_a^b F(x, u, u')dx$$

necessarily satisfies the Euler-Lagrange equation

$$F_u - \frac{d}{dx} F_{u'} = 0$$

with so-called natural boundary conditions

$$F_{u'} = 0$$

in $x = a$ and $x = b$.

Calculus of Variations

Euler-Lagrange Equation in 2-D

$$E(u) = \int_\Omega F(x, y, u, u_x, u_y)dxdy$$

yields the Euler-Lagrange equation

$$F_u - \frac{d}{dx}F_{u_x} - \frac{d}{dy}F_{u_y} = 0$$

with the natural boundary condition

$$\mathbf{n}^T \begin{pmatrix} F_{u_x} \\ F_{u_y} \end{pmatrix} = 0$$

on the rectangular boundary $\partial\Omega$ with normal vector $\mathbf{n}$.
Extensions to higher dimensions are analogous.

Calculus of Variations

Euler-Lagrange Equations for Vector-Valued Functions

$$E(u, v) = \int_a^b F(x, u, v, u', v')dx$$

creates a set of Euler-Lagrange equations:

$$F_u - \frac{d}{dx}F_{u'} = 0$$
$$F_v - \frac{d}{dx}F_{v'} = 0$$

with natural boundary conditions for u and v.
Extensions to vector-valued functions with more components are straightforward.

Lagrange Multipliers

- Sometimes we need to optimise a function with respect to some constraints.
  - Minimise $f(x)$ subject to $x > 0$.
  - Maximise $f(x)$ subject to $g(x) = 0$.
- The method of **Lagrange Multipliers** is an elegant way of optimising functions subject to some constraints.
- The point x for which $\nabla f(x) = 0$ is called the **stationary point** of f.
- Method of Lagrange multipliers finds the stationary points of a function subject to one or more constraints.

Lagrange Multipliers

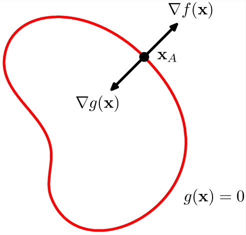
- For a D dimensional vector $\mathbf{x}, g(\mathbf{x}) = 0$ is a $D-1$ dimensional surface in $\mathbf{x}$-space.
- Let $\mathbf{x}$ and $\mathbf{x}+\boldsymbol{\epsilon}$ be two nearby points on the surface $g(\mathbf{x}) = 0$.
- Using Taylor's expansion around $\mathbf{x}$

$$
\begin{aligned}
g(\mathbf{x}+\boldsymbol{\epsilon}) &\approx g(\mathbf{x}) + \boldsymbol{\epsilon}^T \nabla g(\mathbf{x}) \\
\implies \boldsymbol{\epsilon}^T \nabla g(\mathbf{x}) &\approx \mathbf{0}
\end{aligned}
$$

- In the limit $||\boldsymbol{\epsilon}|| \to 0$
  - $\boldsymbol{\epsilon}$ becomes parallel to the constraint surface $g(\mathbf{x}) = 0$, and
  - $\boldsymbol{\epsilon}^T \nabla g(\mathbf{x}) = \mathbf{0}$
- Therefore, $\nabla g(\mathbf{x})$ must be orthogonal to the surface $g(\mathbf{x}) = 0$.

Lagrange Multipliers

- For any surface $g(\mathbf{x}) = 0$, the gradient $\nabla g(\mathbf{x})$ is orthogonal to the surface.
- At any maximiser $\mathbf{x}^*$ of $f(\mathbf{x})$ that also satisfies $g(\mathbf{x}) = 0$, $\nabla f(\mathbf{x})$ must also be orthogonal to the surface $g(\mathbf{x}) = 0$.
  - If $\nabla f(\mathbf{x})$ is orthogonal to $g(\mathbf{x}) = 0$ at $\mathbf{x}^*$, then any movement around $\mathbf{x}^*$ along surface $g(\mathbf{x}) = 0$ is orthogonal to $\nabla f(\mathbf{x})$ and will not increase the value of f.
  - The only way to increase value of f at $\mathbf{x}^*$ is to leave the constraint surface $g(\mathbf{x}) = 0$.



Lagrange Multipliers

- So, at any maximiser $\mathbf{x}^*$, ∇f and ∇g are parallel (or anti-parallel) vectors.
- This can be stated mathematically as

$$\nabla f + \lambda \nabla g = 0$$

  where $\lambda \neq 0$ is the so-called **Lagrange multiplier**.
- This can also be formulated as maximisation of the so-called **Lagrangian function**

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

  with respect to $\mathbf{x}$ and λ.
- Note that this maximisation is unconstrained.

Lagrange Multipliers

At maximiser $\mathbf{x}^*$

$$0 \equiv \nabla L = \nabla f(\mathbf{x}) + \lambda \nabla g(\mathbf{x})$$

which gives $D+1$ equations that the optimal $\mathbf{x}^*$ and λ^* must satisfy

$$\frac{\partial L}{\partial x_1} = 0$$
$$\vdots$$
$$\frac{\partial L}{\partial x_D} = 0$$
$$\frac{\partial L}{\partial \lambda} = 0$$

If only $\mathbf{x}^*$ is required then λ can be eliminated without determining its value (hence λ is also called an **undetermined multiplier**.)

Lagrange Multipliers

Example

Maximise $1 - x_1^2 - x_2^2$ subject to the constraint $x_1 + x_2 = 1$.