CS-567 Machine Learning

Nazar Khan

PUCIT

Lectures 5-8
Oct 27, 29 and Nov 3, 5 2015

Gaussian Distribution

- Known as the queen of distributions.
- Also called the **Normal distribution** since it models the distribution of almost all natural phenomenon.
- For continuous variables.

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

  where μ is the **mean**, σ^2 is the **variance** and σ is the **standard deviation**.
- Reciprocal of variance, $\beta = \frac{1}{\sigma^2}$ is called **precision**.

Gaussian Distribution

- Multivariate form for $D-$ dimensional vector $\mathbf{x}$ of continuous variables

$$\mathcal{N}(\mathbf{x}|\mu, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D|\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

where the $D \times D$ matrix $\boldsymbol{\Sigma}$ is called the **covariance matrix** and $|\boldsymbol{\Sigma}|$ is its determinant.

Independent and Identically Distributed

- Let $\mathcal{D} = (x_1, \ldots, x_N)$ be a set of N random numbers.
- If value of any x_i does not affect the value of any other x_j, then the x_is are said to be **independent**.
- If each x_i follows the same distribution, then the x_is are said to be **identically distributed**.
- Both properties combined are abbreviated as **i.i.d**.
- Assuming the x_is are i.i.d under $\mathcal{N}(\mu, \sigma^2)$

$$p(\mathcal{D}|\mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \sigma^2)$$

- This is known as the **likelihood function** for the Gaussian.

Fitting a Gaussian

- Assuming we have i.i.d data $\mathcal{D} = (x_1, \ldots, x_N)$, how can we find the parameters of the Gaussian distribution that generated it?
- Find the (μ, σ^2) that **maximise the likelihood**. This is known as the **maximum likelihood (ML)** approach.
- Since logarithm is a monotonically increasing function, maximising the log is equivalent to maximising the function.
- Logarithm of the Gaussian
  - is a simpler function, and
  - is numerically superior (consider taking product of very small probabilities versus taking the sum of their logarithms).

Log Likelihood

- Log likelihood of Gaussian becomes

$$\ln p(\mathcal{D}|\mu, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

- Maximising w.r.t μ, we get

$$\mu_{ML} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- Maximising w.r.t σ^2, we get

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2$$

Bias of Maximum Likelihood

- Exercise 1.12
- Since $\mathbb{E}\left[\mu_{ML}\right] = \mu$, ML estimates the mean correctly.
- But since $\mathbb{E}\left[\sigma^2_{ML}\right] = \left(\frac{N-1}{N}\right)\sigma^2$,
  <u>ML underestimates the variance</u> by a factor $\frac{N-1}{N}$.
- This phenomenon is called **bias** and lies at the root of over-fitting.

Polynomial Curve Fitting
A Probabilistic Perspective

- Our earlier treatment was via error minimization.
- Now we take a probabilistic perspective.
- The real goal: make accurate prediction t for new input x given training data $(\mathbf{x}, \mathbf{t})$.
- Prediction implies uncertainty. Therefore, target value can be modelled via a probability distribution.
- We assume that given x, the target variable t has a Gaussian distribution.

$$
\begin{aligned}
p(t|x, \mathbf{w}, \beta) &= \mathcal{N}(t|y(x, \mathbf{w}), \beta^{-1}) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(t - y(x, \mathbf{w}))^2\right\}
\end{aligned} \tag{1}
$$

Polynomial Curve Fitting
A Probabilistic Perspective

- Knowns: Training set $(\mathbf{x}, \mathbf{t})$.
- Unknowns: Parameters $\mathbf{w}$ and β.
- Assuming training data is i.i.d likelihood function becomes

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathbf{w}), \beta^{-1})$$

- Log of likelihood becomes

$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta^{-1} - \frac{N}{2} \ln(2\pi)$$

- Maximization of likelihood w.r.t $\mathbf{w}$ is equivalent to minimization of $\frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2$.

Polynomial Curve Fitting
A Probabilistic Perspective

- **So**, assuming $t \sim \mathcal{N}$, ML estimation leads to sum-of-squared errors minimisation.
- **Equivalently**, minimising sum-of-squared errors implies $t \sim \mathcal{N}$ (*i.e.*, noise was normally distributed).

Polynomial Curve Fitting

A Probabilistic Perspective

- $\mathbf{w}_{ML}$ and β_{ML} yields a probability distribution over the prediction t.

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}_{ML}, \beta_{ML}) = \prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, \mathbf{w}_{ML}), \beta_{ML}^{-1})$$

- The polynomial function $y(x, \mathbf{w}_{ML})$ alone only gives a point estimate of t.

Polynomial Curve Fitting
Bayesian Perspective

- ML estimation of $\mathbf{w}$ maximises the likelihood function $p(\mathbf{t}|\mathbf{x}, \mathbf{w})$ to find the $\underline{\mathbf{w} \text{ for which the observed data is most likely}}$.
- By using a prior $p(\mathbf{w})$, we can employ Bayes' theorem

$$\underbrace{p(\mathbf{w}|\mathbf{x}, \mathbf{t})}_{\text{posterior}} \propto \underbrace{p(\mathbf{t}|\mathbf{x}, \mathbf{w})}_{\text{likelihood}} \underbrace{p(\mathbf{w})}_{\text{prior}}$$

- Now maximise the posterior probability $p(\mathbf{w}|\mathbf{x}, \mathbf{t})$ to find the $\underline{\text{most probable } \mathbf{w} \text{ given the data } (\mathbf{x}, \mathbf{t})}$.
- This technique is called **maximum posterior** or **MAP**.

Polynomial Curve Fitting
Bayesian Perspective

- Let the prior on parameters $\mathbf{w}$ be a zero-mean Gaussian

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) = \left(\frac{\alpha}{2}\right)^{(M+1)/2} \exp\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\}$$

- Negative logarithm of posterior becomes

$$-\ln p(\mathbf{w}|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

which is the same as the *regularized sum-of-squares error* function with $\lambda = \alpha/\beta$.

Polynomial Curve Fitting

Bayesian Perspective

- **So**, assuming $t \sim \mathcal{N}$ **and** $\mathbf{w} \sim \mathcal{N}$, MAP estimation leads to regularized sum-of-squared errors minimisation.
- **Equivalently**, minimising regularized sum-of-squared errors implies $t \sim \mathcal{N}$ **and** $\mathbf{w} \sim \mathcal{N}$ (*i.e.*, noise and the parameters were normally distributed).
- If precision on noise and parameters were α and β respectively, then regularizer $\lambda = \alpha/\beta$.
- MAP estimation allows us to determine optimal α and β whereas regularised-SSE minimisation depends on a user-given λ.

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.

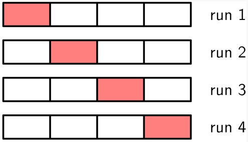


Figure: S-fold cross validation for $S = 4$. Every training is evaluated on the validation set (in red) and these validation set perfromance are averaged over the S training runs.

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.

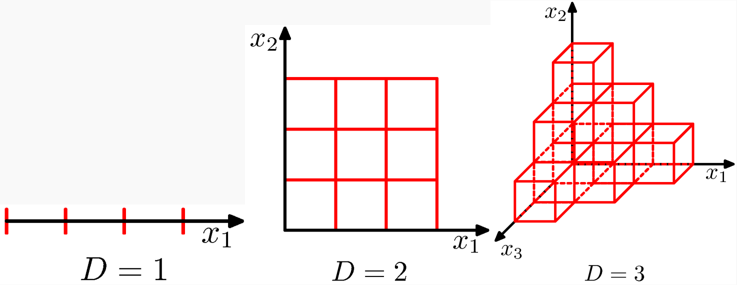


Figure: Curse of Dimensionality: The number of regions of a regular grid grows exponentially with with the dimensionality D of the search space.

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

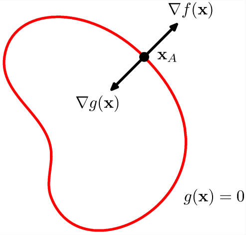
- For a D dimensional vector $\mathbf{x}, g(\mathbf{x}) = 0$ is a $D-1$ dimensional surface in $\mathbf{x}$-space.
- Let $\mathbf{x}$ and $\mathbf{x}+\epsilon$ be two nearby points on the surface $g(\mathbf{x}) = 0$.
- Using Taylor's expansion around $\mathbf{x}$

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

- In the limit $||\epsilon|| \to 0$
  - ϵ becomes parallel to the constraint surface $g(\mathbf{x}) = 0$, and
  - $\epsilon^T \nabla g(\mathbf{x}) = \mathbf{0}$
- 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 $x_1 + x_2 = 1$.