Advanced Math Notes

Yuchen Wang

June 16, 2020

Contents

1 Free parameter

A variable in a mathematical model which cannot be predicted precisely or constrained by the model and must be estimated experimentally or theoretically.

2 Rectifier

2.1 Definition

An activation function defined as the positive part of its argument:

$$f(x) = \max(0, x)$$

Also known as: ramp function
A unit employing the rectifier is also called a **rectified linear unit (ReLU)**

2.2 Softplus (Smooth ReLu)

A smooth approximation to the rectifier is the analytic function

$$f(x) = \log(1 + e^x)$$

Also known as: SmoothReLU
The derivative of softplus is

$$f'(x) = \frac{1}{1 + e^{-x}}$$

(the logistic function)

Notes The logistic function is a smooth approximation of the derivative of the rectifier, the **Heaviside step function**

2.3 LogSumExp: Multivariable Generalization to Softplus

LogSumExp with the first argument set to zero

$$LSE_0^+(x_1, \ldots, x_n) := LSE(0, x_1, \ldots, x_n) = \log(1 + e^{x_1} + \ldots + e^{x_n})$$

Notes The LogSumExp function itself is:

$$LSE(x_1, \ldots, x_n) = \log(e^{x_1} + \ldots + e^{x_n})$$

and its gradient is the softmax.
The softmax with the first argument set to zero is the multivariable generalization of the logistic function.

3 Softmax Function

The softmax function takes an un-normalized vector, and normalizes it into a probability distribution. That is, prior to applying softmax, some vector elements could be negative, or greater than one; and might not sum to 1; but after applying softmax, each element x_i is in the interval $[0, 1]$, and $\sum_i x_i = 1$

$$\sigma : \mathbb{R}^K \to \{\sigma \in \mathbb{R}^K | \sigma_i > 0, \sum_{i=1}^{K} \sigma_i = 1\}$$

$$\sigma(\mathbf{z})_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

for $j = 1, \ldots, K$

4 Cross Entropy

The Cross entropy between two probability distributions p and q over the same underlying set of events measures the average number of bits needed to identify an even drawn from the set if a coding scheme used for the set is optimized for an estimated probability distribution q, rather than the true distribution p.

Discrete distributions

$$H(p, q) = -\sum_{x \in \chi} p(x) \log q(x)$$

Continuous distributions

$$H(p, q) = -\int_{\chi} P(x) \log Q(x)\, dr(x)$$

5 Cross Product in Higher Dimensions

A way of turning 3 vectors in 4-space into a fourth vector, orthogonal to the others, in a trilinear way
Canonical basis of $\mathbb{R}^4 : (e_1, e_2, e_3, e_4)$. If your vectors are $\mathbf{t} = (t_1, t_2, t_3, t_4), \mathbf{u} = (u_1, u_2, u_3, u_4)$ and $\mathbf{v} = (v_1, v_2, v_3, v_4)$, then compute the determinant:

$$\begin{vmatrix} t_1 & t_2 & t_3 & t_4 \\ u_1 & u_2 & u_3 & u_4 \\ v_1 & v_2 & v_3 & v_4 \\ e_1 & e_2 & e_3 & e_4 \end{vmatrix}$$

The cross product of $\mathbf{t}, \mathbf{u}, \mathbf{v}$ is:

$$-e_1 \begin{vmatrix} t_2 & t_3 & t_4 \\ u_2 & u_3 & u_4 \\ v_2 & v_3 & v_4 \end{vmatrix} + e_2 \begin{vmatrix} t_1 & t_3 & t_4 \\ u_1 & u_3 & u_4 \\ v_1 & v_3 & v_4 \end{vmatrix} - e_3 \begin{vmatrix} t_1 & t_2 & t_4 \\ u_1 & u_2 & u_4 \\ v_1 & v_2 & v_4 \end{vmatrix} + e_4 \begin{vmatrix} t_1 & t_2 & t_3 \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_3 \end{vmatrix}$$

6 Gaussian Process

6.1 The Basics

Definition 1 We use a Gaussian process to describe a distribution over functions:

$$\mathbf{f} \sim \mathcal{GP}(m, K)$$

where $m : \chi \to \mathbb{R}$ is the mean function

$$m(\mathbf{x}) = E[f(\mathbf{x})]$$

and $K : \chi^2 \to \mathbb{R}$ is the covariance function

$$K(\mathbf{x}, \mathbf{x}') = E[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$$

Definition 2 For any set S, a Gaussian Process on S is a set of r.v.s $Z_t : t \in S$ s.t. $\forall n \in \mathbb{N}, \forall t_1, \ldots, t_n \in S, (Z_{t_1}, \ldots, Z_{t_n})$ is multi-variate Gaussian.

Theorem: Existence of Gaussian Process For any set S, any mean function $\mu : S \to \mathbb{R}$, and any covariance function $k : S \times S \to \mathbb{R}$, there exists a GP Z_t on S s.t. $E(Z_t) = \mu(t), Cov(Z_s, Z_t) = k(s, t) \, \forall s, t \in S$.

GPs define multivariate Gaussian distributions We have data points $X = [\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T]^T$ and are interested in their function values $\mathbf{f}(X) = (f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n))^T$.
$\mathbf{f}$ is one subset of r.v. and has (prior) joint Gaussian distribution:

$$\mathbf{F}(X) \sim \mathcal{N}(\mathbf{m}(X), K(X, X))$$

Remarks

1. The covariance function $K(\mathbf{x}, \mathbf{x}')$ returns a measure of the similarity of $\mathbf{x}$ and $\mathbf{x}'$ that also encodes how similar $f(\mathbf{x})$ and $f(\mathbf{x}')$ should be.
2. The mean function $m(\mathbf{x})$ encodes a prior expectation of the (unknown) function

Setting the mean function In most cases we simply use

$$E(f(\mathbf{x})) = m(\mathbf{x}) = 0$$

which makes sense especially if we normalize the output to zero mean.

Properties of the covariance function The covariance function $K(\mathbf{x}, \mathbf{x}')$ needs to be a measure of similarity between $\mathbf{x}$ and $\mathbf{x}'$.

1. K needs to be symmetric
$$K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$$
2. K needs to be positive semidefinite (nonnegative definite)
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K(\mathbf{x}, \mathbf{x}') g(\mathbf{x}) g(\mathbf{x}') \, d\mathbf{x} d\mathbf{x}' \geq 0$$
for all $g \in L_2$.

Setting the covariance function

1. Gaussian Kernel:
$$K(r) = \theta_A^2 \exp[-\frac{r^2}{2\theta_L^2}]$$
2. Periodic Covariance Function
$$K(r) = \theta_A^2 \exp[-\frac{\sin^2[(2\pi/\theta_P)r]}{2}]$$
where $r = ||x - x'||$ denotes the Euclidean distance between two indexes.

Hyperparameters θ_A: y-scaling
θ_L: x-scaling (or time scale if the data are time series)
θ_P: period of the covariance functions

6.2 Gaussian Process Regression

Basically equivalent to Bayesian linear regression.
Twist is that using kernel instead of basis functions in order to define the family of functions that you are using for regression.
This allows us to define a very rich family of functions that using basis functions alone could not handle (e.g. mapping into an infinite dimensional space).
In a Gaussian Process regression model, mathematically the same inference as in linear regression can be done.

The model Let $Z \in \mathbb{R}^n \sim N(\mu, K), \varepsilon \in \mathbb{R}^n \sim N(0, \sigma^2 I)$ be independent r.v.s. Let $y = Z + \varepsilon$, so $y \sim N(\mu, K + \sigma^2 I)$.
Define $C = K + \sigma^2 I$.
Let $a = (1, \ldots, l), b = (l+1, \ldots, n)$, so $y = \begin{pmatrix} y_a \\ y_b \end{pmatrix}$, where $y_a = \begin{pmatrix} y_1 \\ \vdots \\ y_l \end{pmatrix}, y_b = \begin{pmatrix} y_{l+1} \\ \vdots \\ y_n \end{pmatrix}$. In addition, $\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, C = \begin{pmatrix} C_{aa} & C_{ab} \\ C_{ba} & C_{bb} \end{pmatrix}, K = \begin{pmatrix} K_{aa} & K_{ab} \\ K_{ba} & K_{bb} \end{pmatrix}$
Then we have $(Y_a | Y_b = y_b) \sim N(m, D)$, where

$$\begin{aligned} m &= \mu_a + C_{ab} C_{bb}^{-1} (y_b - \mu_b) \\ &= \mu_a + K_{ab} (K_{bb} + \sigma^2 I)^{-1} (y_b - \mu_b) \\ D &= C_{aa} - C_{ab} C_{bb}^{-1} C_{ba} \\ &= (K_{aa} + \sigma^2 I) - K_{ab} (K_{bb} + \sigma^2 I)^{-1} K_{ba} \end{aligned}$$

Parameters

1. μ
2. K

Inference

1. Plot the mean function to predict unobserved values (good for visualization)
2. Plot the error curves
3. Choose a loss function and minimize the loss using posterior distribution

Negative Log Marginal Likelihood (NLML) The values of hyperparameters θ may be optimized by minimizing NLML:

$$\begin{aligned} NLML &= -\log p(\mathbf{y}|\mathbf{x},\theta) \\ &= \frac{1}{2}\log|K| + \frac{1}{2}\mathbf{y}^T K^{-1}\mathbf{y} + \frac{n}{2}\log(2\pi) \end{aligned}$$

7 Kronecker Product

A generalization of the outer product from vectors to matrices.

Definition

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix}$$

8 Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

where A and B are events and $P(B) \neq 0$.

9 Maximum a Posteriori Estimation (MAP)

An estimate of an unknown quantity that equals the mode of the posterior distribution.
Can be used to obtain a point estimate of an unobserved quantity on the basis of empirical data.
Closely related to MLE but employs an augmented optimization objective which incorporates a prior distribution (fix the overfitting problem).
Can be seen as a regularization of MLE.

Description Assume we want to estimate an unobserved population parameter θ on the basis of observations x¿ Let f be the sampling distribution of x so that $f(x|\theta)$ is the probability of x when the underlying population parameter is θ.
Now assume that a prior distribution g exists. This allows us to treat θ as a random variable. By Bayes' Theorem, the posterior distribution of θ is

$$f(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int_\Theta f(x|v)g(v)\,dv}$$

where g is the density function of θ, Θ is the domain of g. Then

$$\hat{\theta}_{MAP}(x) = \underset{\theta}{argmax} f(\theta|x)$$

10 Bayesian Linear Regression

Why not use MLE? - overfitting
Why not use MAP? - no representation of uncertainty

Setup

$$D = ((x_1, y_1), \ldots, (x_n, y_n)), x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$$

Model $y_1, \ldots, y_n$ indep given w, $y_i \sim N(w^T x_i, a^{-1}), a > 0$
$w \sim N(0, b^{-1}I), b > 0. (w = (w_1, \ldots, w_d))$
$(a, b := 1/variance$ is called precision)
Assume a, b are known so the only parameter is w.

Posterior Distribution of w It can be shown that

$$P(w|D) = N(w|\mu, \Lambda^{-1})$$

where $\mu = a\Lambda^{-1}X^T y$ and $\Lambda = aX^T X + bI$ where X is the design matrix.

11 Laplace Distribution

Sometimes also called "double exponential distribution", because it can be thought of as two exponential distributions (with an additional location parameter) spliced together back-to-back.

PDF

$$\frac{1}{2b} \exp(-\frac{|x - \mu|}{b})$$

12 Kullback-Leibler Divergence

Also called "relative entropy", is a measure of how one probability distribution is different from a second, reference probability distribution.

Definition For discrete probability distribution P and Q defined on the same probability space, the Kullback-Leibler divergence between P and Q is defined to be

$$D_{KL}(P|Q) = -\sum_{x \in \chi} P(x) \log(\frac{Q(x)}{P(x)})$$

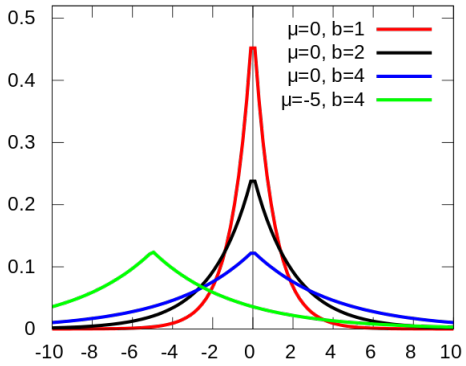

Figure 1: PDF of Laplace Distribution

For distributions P and Q of a continuous random variable, the Kullback-Leiber divergence is defined to be the integral

$$D_{KL}(P|Q) = \int_{-\infty}^{\infty} p(x) \log(\frac{p(x)}{q(x)})\, dx$$

where p and q denote the probability densities of P and Q.

13 Power Set

Given a set S, the power set of S, is the set of all subsets of S.
Notation: 2^S or $\mathcal{P}(S)$

14 the Binomial Theorem

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$$

15 Markov Process

15.1 Introduction

Definition 15.1.1 Stochastic Process A stochastic process is a sequence of events in which the outcome at any stage depends on some probability.

Definition 15.1.2 Markov Process A Markov process is a stochastic process with the following properties:

1. The number of possible outcomes or states is finite
2. The outcome at any stage depends only on the outcome of the previous stage
3. The probabilities are constant over time

Definition 15.1.3 Markov Chain If $\mathbf{x}_0$ is a vector which represents the initial state of a system, then there is a matrix M such that the state of the system after one iteration is given by the vector $M\mathbf{x}_0$. Thus we get a chain of state vectors:

$$\mathbf{x}_0, M\mathbf{x}_0, M^2\mathbf{x}_0, \ldots$$

where the state of the system after n iterations is given by $M^n\mathbf{x}_0$. Such a chain is called a Markov chain and the matrix M is called a transition matrix.
The state vectors can be of one of two types: an absolute vector or a probability vector.
An absolute vector is a vector whose entries give the actual number of objects in a given state.
A probability vector is a vector where the entries give the percentage (or probability) of objects in a given state. Note that the entries of a probability vector add up to 1.

Theorem 15.1.4 Let M be the transition matrix of a Markov process such that M^k has only positive entries for some k. Then there exists a unique probability vector $\mathbf{x}_s$ such that

$$M\mathbf{x}_s = \mathbf{x}_s$$

Moreover

$$\lim_{k\to\infty} M^k\mathbf{x}_0 = \mathbf{x}_s$$

for any initial state probability vector $\mathbf{x}_0$.

15.2 The Transition Matrix and its Steady-state Vector

The transition matrix of an n-state Markov process is an $n \times n$ matrix M where the i, j entry of M represents the probability that an object in state j transitions into state i, that is if $M = (\mathbf{m}_{ij})$ and the states are $S_1, S_2, \ldots, S_n$ then $\mathbf{m}_{ij}$ is the probability that an object in state S_j transitions to state S_i.

16 Jensen's Inequality (Probability Theory)

If X is a random variable and φ is a convex function, then

$$\varphi(E[X]) \leq E[\varphi(X)]$$

17 Weight Decay

An example of a regularization method.
One way to penalize complexity would be to add all weights to the loss function.
If we have the cost function $E(\mathbf{w})$, then change the cost function to

$$\tilde{E}(\mathbf{w}) = E(\mathbf{w}) + \frac{\lambda}{2}\mathbf{w}^2$$

Where λ is the regularization parameter.
Applying gradient descent to this new cost function we obtain:

$$w_i \leftarrow w_i - \eta\frac{\partial E}{\partial w_i} - \eta\lambda w_i$$

18 De Moivre's Formula

For any real number x and integer n, it holds that

$$(\cos(x) + i\sin(x))^n = \cos(nx) + i\sin(nx)$$

where i is the imaginary unit ($i^2 = -1$).

Relation to Euler's formula De Moivre's formula is a precursor to Euler's formula

$$e^{ix} = \cos x + i\sin x$$

One can derive de Moivre's formula using Euler's formula and the exponential law for integer powers

$$\left(e^{ix}\right)^n = e^{inx}$$

since Euler's formula implies that the left side is equal to $(\cos x + i\sin x)^n$ while the right side is equal to $e^{inx} = \cos(nx) + i\sin(nx)$.

19 Bayesian Optimization

We are interested in finding the minimum of a function $f(x)$ on some bounded set $\chi¿$ which we will take to be a subset of $\mathbb{R}^D$. Bayesian optimization constructs a probabilistic model for $f(x)$ and then exploits this model to make decisions about where in χ to next evaluate the function, while integrating out uncertainty.
There are two major choices made when performing Bayesian optimization

1. Select a prior (assumptions) over functions being optimized;
2. Choose an **acquisition function**, which is used to construct a utility function from the model posterior, allowing us to determine the next point to evaluate.

19.1 Gaussian Processes (GP) as the prior

We will take GP to be of the form $f : \chi \to \mathbb{R}$. The GP is defined by the property that any finite set of N points $\{\mathbf{x} \in \chi\}_{n=1}^N$ induces a multivariate Gaussian distribution on $\mathbb{R}^N$. The nth of these points is taken to be the function value $f(\mathbf{x}_n)$.
We assume that the function $f(\mathbf{x})$ is drawn from a Gaussian process prior and that our observation are of the form $\{\mathbf{x}_n, y_n\}_{n=1}^N$, where $y_n \sim \mathcal{N}(f(\mathbf{x}_n), v)$ and v is the variance of noice introduced into the function observations. This prior and data induce a posterior over functions.
The acquisition function, which we denote by $a : \chi \to \mathbb{R}^+$, determines what point in χ should be evaluated next via a proxy optimization $\mathbf{x}next = \text{argmax}_{\mathbf{x}}\, a(\mathbf{x})$, where several different functions have been proposed. In general, these acquisition functions depend on the previous observations and GP hyperparameters. We denote this dependence as $a(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta)$.
Define γ to be the function that normalizes $f(\mathbf{x}_{best})$:

$$\gamma(\mathbf{x}) = \frac{f(\mathbf{x}_{best}) - \mu(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta)}{\sigma(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta)}$$

Probability of Improvement (PI) One strategy is to maximize the probability of improving over the best current value.

$$a_{PI}(\mathbf{x}) = P(f(\mathbf{x}) < \gamma(\mathbf{x}))$$

Under GP,

$$a_{PI}(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta) = \Phi(\gamma(x))$$

Expected Improvement Alternatively, one could choose to maximize the expected improvement over the best current value.

$$a_{EI}(\mathbf{x}) = \mathbb{E}[\max(\gamma(\mathbf{x}) - f(\mathbf{x}), 0))]$$

(The idea: if the new value is much better, we win by a lot; if it's much worse, we haven't lost anything)
Under GP,

$$a_{EI}(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta) = \sigma(\mathbf{x}; \{\mathbf{x}_n, y_n\}, \theta)(\gamma(\mathbf{x})\Phi(\gamma(\mathbf{x})) + (\gamma(\mathbf{x}); 0, 1))$$

19.2 Jointly Continuity and Separately Continuity

Definition 19.1 (jointly continuity and separately continuity)**.** If X, Y and Z are topological spaces and $f : X \times Y \to Z$ is a function then we say that f is **jointly continuous** at $(x_0, y_0) \in X \times Y$ if for each neighbourhood W of $f(x_0, y_0)$ there exists a product of open sets $U \times V \subseteq X \times Y$ containing (x_0, y_0) such that $f(U \times V) \subseteq W$.

We say that f is **separately continuous** on $X \times Y$ if for each $x_0 \in X$ and $y_0 \in Y$ the functions $y \mapsto f(x_0, y)$ and $x \mapsto f(x, y_0)$ are both continuous on Y and X respectively. If the range space Z is a metric space, with metric d, and ε is a positive number then we say that f is ε**-jointly continuous**xw at $(x_0, y_0) \in X \times Y$ if there exists a product of open sets $U \times V \subseteq X \times Y$ containing $(x_0, y_0) \in X \times Y$ such that d-diam $f(U \times V) \leq \varepsilon$.