Advanced Math Notes

Yuchen Wang

May 26, 2019

Contents

1 Free parameter **2**

2 Rectifier **2**
2.1 Definition 2
2.2 Softplus 2
2.3 Multivariable Generalization to Softplus 2

3 Softmax Function **3**

4 Cross Entropy **3**

5 Cross Product in Higher Dimensions **4**

6 Gaussian Process **4**
6.1 The Basics 4
6.2 Gaussian Process Regression 6

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

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 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 $y \sim N(\mu, K + \sigma^2 I)$.

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 $(Y_a | Y_b = y_b) \sim (m, D)$, 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:

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