Dự án Cuối kỳ

Môn: Nhập môn Học Máy

Họ và Tên: Đoàn Nguyễn Minh Phương

MSSV: 51900182 – 19050201

**Bài 1:**

Trình bày một bài nghiên cứu, đánh giá của em về các vấn đề sau:

1. Tìm hiểu, so sánh các phương pháp Optimizer trong huấn luyện mô hình học máy;

|  |  |  |  |
| --- | --- | --- | --- |
| Phương pháp | Mô tả | Ưu điểm | Nhược điểm |
| Batch Gradient Descent (BGD) | Toàn bộ dữ liệu được sử dụng để tính độ dốc của hàng chi phí đối với các tham số trong một bước. Nó tính toán độ dốc trung bình của toàn bộ dữ liệu | Hội tụ giá trị tối ưu toàn bộ cục đối với hàm lối. | Tính toán tốn kém cho các bộ dữ liệu lớn. Yêu cầu toàn bộ bộd ữ liệu trong bộ nhớ |
| Stochastic Gradient Descent (SGD) | Trong độ dốc ngẫu nhiên, một điểm dữ liệu được chọn ngẫu nhiên được sử dụng để tính độ dốc và cập nhật các tham số trong mỗi lần lặp. Nó cung cấp cập nhật nhanh hơn nhưng giới thiệu nhiều phương sai trong các cập nhật tham số. | Hiệu quả về mặt tính toán, đặc biệt là đối với các bộ dữ liệu lớn. | Phương sai cao trong các cập nhật, có thể dao động xung quanh giá trị tối thiểu. |
| Mini-Batch Gradient Descent | Độ dốc nhỏ hàng loạt là sự thoả hiệp giữa độ dốc hàng loạt và độ dốc ngẫu nhiên. Nó sử dụng một tập con nhỏ được chọn ngẫu nhiên (độ dốc nhỏ) của bộ dữ liệu để tính độ dốc và cập nhật các tham số. | Cân bằng giữa hiệu suất và nhiễu; phù hợp với xử lý song song. | Vẫn giới thiệu một số phương sai |
| Adam (Adaptive Moment Estimation) | Adam kết hợp những ưu điểm của cả động lượng và RMSprop. Nó duy trì hai đường trung bình động cho mỗi tham số – khoảnh khắc đầu tiên (trung bình) và khoảnh khắc thứ hai (phương sai không tập trung). Nó điều chỉnh tốc độ học tập cho từng tham số riêng lẻ | Tốc độ học tập thích ứng, phù hợp với nhiều vấn đề khác nhau. | Có thể nhạy cảm với siêu tham số; có thể không hoạt động tốt trên tất cả các loại dữ liệu. |
| RMSprop (Root Mean Square Propagation) | AdaDelta là phần mở rộng của RMSprop nhằm loại bỏ nhu cầu thiết lập tốc độ học theo cách thủ công. Nó sử dụng tỷ lệ bình phương trung bình gốc của tham số cập nhật với bình phương trung bình gốc của gradient | Điều chỉnh tốc độ học tập; loại bỏ sự cần thiết phải thiết lập tốc độ học tập theo cách thủ công | Nhạy cảm với siêu tham số. |
| AdaGrad | AdaGrad điều chỉnh tốc độ học của tất cả các tham số mô hình dựa trên độ dốc lịch sử của chúng. Nó thực hiện cập nhật lớn hơn cho các tham số không thường xuyên và cập nhật nhỏ hơn cho các tham số thường xuyên | Tỷ lệ học tập thích ứng; rất phù hợp cho dữ liệu thưa thớt | Tốc độ học tập có thể trở nên quá nhỏ đối với các tham số thường xuyên xảy ra |
| AdaDelta | AdaDelta là phần mở rộng của RMSprop nhằm loại bỏ nhu cầu thiết lập tốc độ học theo cách thủ công. Nó sử dụng tỷ lệ bình phương trung bình gốc của tham số cập nhật với bình phương trung bình gốc của gradient | Điều chỉnh tốc độ học tập; loại bỏ sự cần thiết phải thiết lập tốc độ học tập theo cách thủ công | Nhạy cảm với siêu tham số |

1. Tìm hiểu về Continual Learning và Test Production khi xây dựng một giải pháp học máy để giải quyết một bài toán nào đó.

**Continuous Learning :**

Continuous learning trong bối cảnh học máy đề cập đến việc thực hành cập nhật và cải tiến mô hình theo thời gian khi có dữ liệu mới. Điều này đặc biệt quan trọng trong môi trường động, nơi các mẫu cơ bản trong dữ liệu có thể thay đổi và mô hình cần phải thích ứng với những thay đổi này. Học tập liên tục bao gồm một số yếu tố chính:

**Online Learning**: Trong Online Learning, mô hình được cập nhật dần dần khi có điểm dữ liệu mới. Điều này trái ngược với phương pháp học theo lô, trong đó mô hình được đào tạo trên toàn bộ tập dữ liệu. Học trực tuyến cho phép mô hình thích ứng nhanh chóng với những thay đổi và có thể sử dụng tài nguyên hiệu quả hơn.

**Giám sát mô hình**: Trong Online Learning, bao gồm việc giám sát hiệu suất của mô hình được triển khai theo thời gian. Điều này bao gồm theo dõi các số liệu hiệu suất chính, xác định bất kỳ sự suy giảm nào về độ chính xác của mô hình và hiểu tác động của các dự đoán mô hình đối với doanh nghiệp hoặc ứng dụng.

**Đào tạo lại**: Định kỳ đào tạo lại mô hình với dữ liệu mới là điều cần thiết để duy trì độ chính xác của mô hình. Tần suất đào tạo lại phụ thuộc vào bản chất của vấn đề và tốc độ phân phối dữ liệu cơ bản phát triển như thế nào.

**Vòng lặp**: Việc thiết lập các vòng phản hồi là rất quan trọng cho việc học tập liên tục. Phản hồi từ người dùng, chuyên gia miền và số liệu hiệu suất hệ thống có thể cung cấp thông tin chi tiết có giá trị về hành vi của mô hình và hướng dẫn cải tiến.

**Thuật toán thích ứng**: Việc sử dụng các thuật toán có thể thích ứng với những thay đổi trong phân phối dữ liệu là điều quan trọng. Một số thuật toán học máy, chẳng hạn như thuật toán học trực tuyến, được thiết kế để điều chỉnh các tham số của chúng khi gặp dữ liệu mới.

**Test Production:**

Test Production đề cập đến thực tiễn tạo ra một chiến lược thử nghiệm toàn diện cho các mô hình học máy và triển khai các thử nghiệm này vào môi trường sản xuất. Mục tiêu là để đảm bảo rằng mô hình đã triển khai hoạt động một cách đáng tin cậy và chính xác, đồng thời nó sẽ tiếp tục hoạt động như vậy theo thời gian. Các yếu tố chính của sản xuất thử nghiệm bao gồm:

**Unit Testing**: Kiểm tra các thành phần riêng lẻ của hệ thống máy học một cách riêng biệt. Điều này bao gồm các bước thử nghiệm tiền xử lý dữ liệu, kỹ thuật tính năng và đào tạo mô hình để đảm bảo chúng hoạt động chính xác.

**Integration Testing**: Đánh giá cách các thành phần khác nhau của hệ thống máy học phối hợp với nhau. Điều này liên quan đến việc kiểm tra quy trình làm việc từ đầu đến cuối, từ đầu vào dữ liệu đến dự đoán mô hình.

**Số liệu đánh giá mô hình**: Xác định và theo dõi các số liệu đánh giá phù hợp cho hiệu suất của mô hình. Các số liệu phổ biến bao gồm độ chính xác, độ chính xác, khả năng thu hồi, điểm F1 và diện tích dưới đường cong đặc tính vận hành máy thu (AUC-ROC).

**Kiểm tra độ ổn định**: Đánh giá độ ổn định của mô hình trong các điều kiện khác nhau, chẳng hạn như những thay đổi trong phân phối dữ liệu, các biến thể của tính năng đầu vào và các sự kiện không mong muốn.

Kiểm tra hiệu suất: Đánh giá hiệu suất tính toán của mô hình, bao gồm tốc độ suy luận và mức sử dụng tài nguyên, để đảm bảo mô hình đáp ứng yêu cầu sản xuất.

**Thử nghiệm A/B:** Tiến hành thử nghiệm A/B trong môi trường sản xuất để so sánh hiệu suất của các phiên bản mô hình khác nhau. Điều này giúp xác định các cải tiến và các vấn đề tiềm ẩn.

**Giám sát và cảnh báo**: Thực hiện giám sát liên tục mô hình đã triển khai để phát hiện các điểm bất thường, suy thoái mô hình hoặc sai lệch so với hành vi dự kiến. Thiết lập hệ thống cảnh báo để thông báo khi có vấn đề phát sinh.

Bằng cách kết hợp học tập liên tục với thực tiễn sản xuất thử nghiệm hiệu quả, các giải pháp học máy có thể vẫn mạnh mẽ, thích ứng và phù hợp với nhu cầu ngày càng tăng của vấn đề mà chúng hướng tới giải quyết. Những thực tiễn này góp phần vào sự thành công lâu dài và độ tin cậy của hệ thống máy học trong các ứng dụng trong thế giới thực.