

ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ

ПРЕДВАРИТЕЛЬНАЯ ПРОГРАММА КУРСА

Сергей Николенко

ПЛАН ЛЕКЦИЙ

Это предварительный план обзорного годового курса. Основным содержанием курса планируется сделать байесовские методы.

- (1) Введение в машинное обучение. История. Постановка задач: обучение с учителем, без учителя, с подкреплением. Статистические и байесовские методы обучения, примеры. Что такое байесовский вывод, основные понятия и термины.
- (2) Простейшие методы обучения с учителем: метод наименьших квадратов, метод ближайших соседей. Регуляризация. Проклятие размерности.
- (3) Классификаторы. Статистическая теория принятия решений.
- (4) Линейная регрессия: регрессия с ядерными функциями, гребневая регрессия, лассо.
- (5) Максимальное правдоподобие и наименьшие квадраты. Bias-variance decomposition. Байесовский подход к линейной регрессии.
- (6) Линейные методы классификации: линейный дискриминантный анализ. Байесовский подход к классификации.
- (7) Оценка и выбор моделей. Байесовский подход: информационные критерии. Размерность Вапника-Червоненкиса. Кроссвалидация. Bootstrapping и bagging.
- (8) Сопряжённые априорные распределения, примеры.
- (9) Кластеризация. Комбинаторные алгоритмы. Алгоритм k-средних. Оценка качества кластеризации.
- (10) Обучение параметров смеси распределений: алгоритм ЕМ. Байесовская кластеризация.
- (11) Марковские цепи: основные понятия. Скрытые марковские модели. Обучение скрытых марковских моделей, алгоритм Баума-Уэлша.
- (12) Графические вероятностные модели: направленные и ненаправленные графы, d-разделимость. Фактор-граф. Задача маргинализации. Алгоритм передачи сообщений.
- (13) Приближённый вывод на графических моделях. Вариационные приближения. Expectation Propagation.
- (14) Маргинализация методами Монте-Карло. МСМС-методы, алгоритм Метрополиса-Гастингса. Сэмплирование по Гиббсу, другие методы сэмплирования.
- (15) Max-margin классификаторы. Метод опорных векторов (SVM). Kernel trick: нелинейные разделяющие поверхности.
- (16) Sparse vector machines. Байесовский подход к SVM: Relevance Vector Machines (RVM).
- (17) Искусственные нейронные сети. Алгоритм обратной пропации ошибок. Байесовский подход к нейронным сетям, оптимизация гиперпараметров.
- (18) Case study: категоризация текстов. Наивный байесовский классификатор. Latent Dirichlet allocation, вывод в модели LDA.
- (19) Бустинг. AdaBoost. Градиентный бустинг. AnyBoost.
- (20) Case study: задача ранжирования и информационного поиска. LambdaRank. Бустинг алгоритмов ранжирования.
- (21) Case study: рекомендательные системы. Онлайн- и оффлайн-системы. Следование за трендами, модель DGP. Коллаборативная фильтрация и модель SVD++.

- (22) Case study: модели поведения пользователя (click models) и рейтинг-системы.
- (23) Case study: разреженное представление сигналов (compressive sensing).
- (24) Обучение с подкреплением. Обучение многоруких бандитов. Алгоритм UCB и его вариации. Бандиты с большим числом рук (many-armed bandits).
- (25) Обучение в известной модели. Марковские процессы принятия решений. Метод временных разностей (TD-обучение). $TD(\lambda)$, Sarsa(λ), $Q(\lambda)$.
- (26) Case study: примеры обучения с подкреплением. Алгоритм TD-Gammon. Другие примеры применений.
- (27) Визуализация многомерных данных. SOM, карты Кохонена.
- (28) Деревья принятия решений. Энтропия, обучение деревьев, алгоритм ID3. Борьба с оверфиттингом, обрезание (pruning), алгоритм C4.5. Альтернирующие деревья и бустинг.
- (29) Деревья принятия решений и бэггинг: случайные леса, случайный наивный Байес, случайная мультиномиальная логистическая регрессия.
- (30) Активное обучение: как выбрать следующую точку для тестирования.

ПРЕДВАРИТЕЛЬНЫЙ ПЛАН СЕМИНАРОВ

План только на первый семестр.

- (1) Вспоминаем теорию вероятностей и статистику. Основные понятия теории вероятностей. Математическое ожидание и дисперсия. Основные распределения вероятностей: дискретные (биномиальное, мультиномиальное) и непрерывные (нормальное, бета-распределение, распределение Пуассона).
- (2) Решение задач по теории вероятностей и математической статистике (уверенное владение базовыми вещами очень важно для основного курса, поэтому планируется уделить этому значительное время).
- (3) Практика машинного обучения: изучаем R и Octave.
- (4) Различные методы линейной регрессии. Практическое задание: реализация линейной регрессии, сравнение методов.
- (5) Линейная классификация на практике. Логистическая регрессия.
- (6) Кластеризация на практике: реализация алгоритма k-средних, алгоритма EM для случая байесовской кластеризации.
- (7) Марковские цепи: решение задач по теме. Реализация алгоритмов обучения скрытых марковских моделей.
- (8) Графические модели и библиотека Infer.NET. Реализация графических моделей в Infer.NET.
- (9) В течение всего семестра – раздача студентам проектов, связанных с анализом данных и машинным обучением (вероятно, в небольших группах; возможно, это будет только для студентов-специалистов: CS и SE Data Mining). Задача – проанализировать тот или иной набор данных известными им методами, провести исследование от начала до конца, сравнить разные методы, сделать выводы о применимости тех или иных методов.