بسم الله الرحمن الرحیم

روش‌های یادگیری ماشین آماری چندمتغیره در پیش‌بینی ژنومی

روش‌های یادگیری ماشین آماری چندمتغیره در پیش‌بینی ژنومی

اُسوال آنتونیو مونته‌سینوس لوپز
آبلاردو مونته‌سینوس لوپز
خوزه کروسا

ترجمهٔ
دکتر عباس سعیدی
دکتر زهره حاجی‌برات
دکتر زهرا حاجی‌برات

۱۴۰۵

۹۴۰
مرکز چاپ و انتشارات دانشگاه شهید بهشتی

روش‌های یادگیری ماشین آماری چندمتغیره در پیش‌بینی ژنومی
اسوال آنتونیو مونته‌سینوس لوپز، آبلاردو مونته‌سینوس لوپز، خوزه کروسا
Osval Antonio Montesinos López, Abelardo Montesinos López, *Multivariate Statistical Machine Learning Methods for Genomic Prediction,* Springer, 2020.
ترجمهٔ دکتر عباس سعیدی، دکتر زهره حاجی‌برات، دکتر زهرا حاجی‌برات

ویراستار: ندا نوری
طراح جلد: امیرشاهرخ فریوسفی
حروف‌نگار و صفحه‌آرا: امیرحسین حیدری
ناظر چاپ: نوید سیفان
چاپ اول: ۱۴۰۵
شمارگان: ۵۰۰
قیمت: ۲۸۸۵۰.۰۰۰ ریال

کلیهٔ حقوق برای دانشگاه شهید بهشتی محفوظ است.
«صحت مطالب کتاب به‌عهدهٔ مترجمان است»

سرشناسه: مونته‌سینوس لوپس، اسوال آنتونیو؛ Montesinos López, Osval Antonio
عنوان و نام پدیدآور: روش‌های یادگیری ماشین آماری چندمتغیره در پیش‌بینی ژنومی/ اسوال آنتونیو مونته‌سینوس لوپز، آبلاردو مونته‌سینوس لوپز، خوزه کروسا؛ ترجمه عباس سعیدی، زهره حاجی‌برات، زهرا حاجی‌برات؛ ویراستار ندا نوری.
مشخصات نشر: تهران: دانشگاه شهید بهشتی، مرکز چاپ و انتشارات، ۱۴۰۵.
مشخصات ظاهری: بیست و هفت، ۷۹۳ ص.: مصور، جدول.
فروست: مرکز چاپ و انتشارات دانشگاه شهید بهشتی؛ ۹۴۰.
شابک: ۹۷۸-۹۶۴-۴۵۷-۶۸۰-۵
یادداشت: عنوان اصلی: Multivariate statistical machine learning methods for genomic prediction,2022.
یادداشت: واژه‌نامه؛ کتابنامه؛ نمایه.
موضوع: فراگیری ماشینی، Machine learning؛ یادگیری عمیق (فراگیری ماشینی)، Deep learning (Machine learning)؛ ژنوم، Genomes
شناسنامه افزوده: مونته‌سینوس لوپس، آبلاردو؛ Montesinos López, Abelardo
شناسه افزوده: کروسا، خوسه؛ Crossa, José
شناسه افزوده: سعیدی، عباس، ۱۳۴۰-، مترجم
شناسه افزوده: حاجی‌برات، زهره، ۱۳۶۷-، مترجم؛ حاجی‌برات، زهرا، ۱۳۶۷-، مترجم
شناسه افزوده: دانشگاه شهید بهشتی، مرکز چاپ و انتشارات
شناسه افزوده: Shahid Beheshti University. Printing & Publishing Center
رده‌بندی کنگره: Q۳۲۵/۵
رده‌بندی دیویی: ۰۰۶/۳۱
شماره کتابشناسی ملی: ۱۰۶۰۱۳۳۳

کد ناشر ۱۰۰۱۷۳۴
unipress.sbu.ac.ir، press.sbu.ac.ir
unipress@mail.sbu.ac.ir

فهرست مطالب

پیشگفتار مترجمان ..... سیزده
سخن آغازین ..... هفده
پیشگفتار نویسندگان ..... بیست‌وسه
فصل اول عناصر رایج انتخاب ژنومی و یادگیری آماری ۱
۱.۱. داده‌ها به‌منزلهٔ سلاحی قدرتمند ۱
۲.۱. انتخاب ژنومی ۵
۱.۲.۱. مفاهیم انتخاب ژنومی ۵
۲.۲.۱. چرا یادگیری ماشین آماری عنصر مهمی در انتخاب ژنومی است؟ ۸
۳.۱. مبانی مدل‌سازی ۱۱
۱.۳.۱. مدل یادگیری ماشین آماری چیست؟ ۱۱
۲.۳.۱. دو شیوهٔ ساخت مدل: پیش‌بینی در مقابل استنباط ۱۲
۳.۳.۱. انواع مدل‌های یادگیری ماشین آماری و تأثیرات آن‌ها ۱۵
۱.۳.۳.۱. انواع مدل‌های یادگیری ماشین آماری ۱۵
۲.۳.۳.۱. تأثیرات مدل ۱۷
۴.۱. بررسی جبر ماتریس‌ها ۲۲
۵.۱. انواع داده‌های آماری ۲۹
۱.۵.۱. انواع داده‌ها ۲۹
۲.۵.۱. انواع داده‌های چندمتغیره ۳۴
۶.۱. انواع یادگیری ۳۴
۱.۶.۱. تعریف یادگیری تحت نظارت و نمونه‌های آن ۳۴
۲.۶.۱. تعریف‌ها و نمونه‌های یادگیری بدون نظارت ۳۸
۳.۶.۱. تعریف و نمونه‌های یادگیری نیمه‌نظارتی ۳۹
منابع فصل اول ۴۰

فصل دوم ابزارهای پیش‌پردازش برای آماده‌سازی داده‌ها ۴۳
۱.۲. تأثیرات ثابت یا تصادفی ۴۳
۲.۲. BLUEها و BLUPها ۴۴
تعریف تابع تخمین‌پذیر ۴۵
توابع پیش‌بینی‌پذیر ۵۰
۳.۲. پالایش نشانگر ۵۴
نشانگرهای ژنتیکی ۵۵
۴.۲. روش‌های محاسبهٔ ماتریس رابطهٔ ژنومی ۶۱
۵.۲. ارزش‌های اصلاح ژنومی و برآورد آن‌ها ۶۴
۶.۲. روش‌های نرمال‌سازی ۷۰
۷.۲. پیشنهادهای معمول برای حذف یا افزودن ورودی‌ها ۷۱
۸.۲. تحلیل مؤلفهٔ اصلی به‌منزلهٔ روشی برای فشرده‌سازی ۷۷
منابع فصل دوم ۸۵
فصل سوم عوامل برای ساخت مدل‌های یادگیری ماشین آماری تحت نظارت ۸۷
۱.۳. تعریف مدل رگرسیون خطی چندگانه ۸۷
۲.۳. برازش یک مدل رگرسیون خطی چندگانه ازطریق روش کمینه‌سازی مربعات خطا (OLS) ۸۷
۳.۳. برازش مدل رگرسیون خطی چندگانه ازطریق روش حداکثر درست‌نمایی (ML) ۹۱
۴.۳. برازش مدل رگرسیون خطی چندگانه ازطریق روش کاهش گرادیان ۹۲
۵.۳. مزایا و معایب مدل‌های رگرسیون خطی استاندارد (کمینه‌سازی مربعات خطا و MLR) ۹۷
۶.۳. مدل رگرسیون خطی چندگانهٔ منظم‌سازی‌شده ۹۸
۱.۶.۳. رگرسیون ریج ۹۸
۲.۶.۳. رگرسیون لاسو ۱۱۲
۷.۳. رگرسیون لجستیک ۱۱۸
۱.۷.۳. رگرسیون ریج لجستیک ۱۲۱
۲.۷.۳. رگرسیون لجستیک لاسو ۱۲۲
منابع فصل سوم ۱۲۹
فصل چهارم بیش‌برازش، تنظیم مدل و ارزیابی پیش‌بینی تابع ۱۳۱
۱.۴. مشکل بیش‌برازش و کم‌برازش ۱۳۱
۲.۴. تعادل میان دقت پیش‌بینی و تفسیرپذیری مدل ۱۳۴
۳.۴. اعتبارسنجی متقابل ۱۳۸
۱.۳.۴. رویکرد مجموعهٔ نگهدارندهٔ واحد ۱۳۹
۲.۳.۴. اعتبارسنجی متقابل k- فولد ۱۳۹
۳.۳.۴. اعتبارسنجی متقابل با حذف یک نمونه ۱۴۱

۴.۳.۴. اعتبارسنجی متقابل با حذف m نمونه ۱۴۲
۵.۳.۴. اعتبارسنجی متقابل تصادفی ۱۴۲
۶.۳.۴. اعتبارسنجی متقابل با حذف یک گروه ۱۴۳
۷.۳.۴. اعتبارسنجی متقابل بازنمونه‌گیری ۱۴۴
۸.۳.۴. اعتبارسنجی متقابل بلوک ناقص ۱۴۵
۹.۳.۴. اعتبارسنجی متقابل تصادفی با استفاده از بلوک‌ها ۱۴۶
۱۰.۳.۴. سایر گزینه‌ها و دیدگاه‌های رایج دربارهٔ اعتبارسنجی متقابل ۱۴۸
۴.۴. تنظیم مدل ۱۵۰
۱.۴.۴. چرا تنظیم مدل مهم است؟ ۱۵۲
۲.۴.۴. روش‌های تنظیم ابرپارامترها (جست‌وجوی شبکه‌ای، جست‌وجوی تصادفی و دیگر شیوه‌ها) ۱۵۳
۵.۴. معیارهای ارزیابی عملکرد پیش‌بینی ۱۵۵
۱.۵.۴. اندازه‌گیری‌های کمّی عملکرد پیش‌بینی ۱۵۶
۲.۵.۴. معیارهای پیش‌بینی کارآیی باینری و ترتیبی ۱۵۹
۳.۵.۴. شمارش اندازه‌گیری‌های عملکرد پیش‌بینی ۱۶۶
منابع فصل چهارم ۱۶۸
فصل پنجم مدل‌های خطی مختلط ۱۷۱
۱.۵. کلیات مدل‌های خطی مختلط ۱۷۱
۲.۵. برآورد مدل خطی مختلط ۱۷۲
۱.۲.۵. برآورد حداکثر درست‌نمایی ۱۷۲
۱.۱.۲.۵. الگوریتم بیشینه‌سازی مورد انتظار ۱۷۴
۲.۱.۲.۵. REML ۱۷۶
۳.۱.۲.۵. BLUPها ۱۷۸
۳.۵. مدل‌های خطی مختلط در پیش‌بینی ژنومی ۱۷۹
۴.۵. مثال‌های گویای مدل‌های خطی مختلط تک‌متغیره ۱۸۰
۵.۵. مدل‌های خطی مختلط ژنومی چندویژگی ۱۸۵
۶.۵. دیدگاه‌های نهایی ۱۹۱
منابع فصل پنجم ۲۰۴
فصل ششم رگرسیون خطی ژنومی بیزی ۲۰۷
۱.۶. قضیهٔ بیزی و رگرسیون خطی بیزی ۲۰۷
۲.۶. رگرسیون ریج مبتنی‌بر ژنوم بیزی ۲۰۸
۳.۶. مدل ژنومی بیزی GBLUP ۲۱۲
۴.۶. مدل پیش‌بینی ژنومی BayesA ۲۱۵
۵.۶. مدل‌های BayesB و BayesC با پیش‌بینی ژنومی ۲۱۷

۶.۶. مدل بیزی لاسو با پیش‌بینی ژنومی ۲۲۲
۷.۶. پیش‌بینی‌کنندهٔ گسترده در رگرسیون ژنومی بیزی مدل‌ها ۲۲۵
۸.۶. مدل رگرسیون خطی ژنومی چندویژگی بیزی ۲۲۹
۱.۸.۶. مدل خطی چندویژگی ژنومی ۲۳۰
۹.۶. مدل چندویژگی و چندمحیطی ژنومی بیزی (BMTME) ۲۳۵
منابع فصل ششم ۲۴۹
فصل هفتم مدل‌های پیش‌بینی بیزی و کلاسیک برای دسته‌بندی و شمارش داده‌ها ۲۵۱
۱.۷. مقدمه ۲۵۱
۲.۷. مدل رگرسیون ترتیبی بیزی ۲۵۱
۱.۲.۷. مثال‌های گویا ۲۵۹
۳.۷. رگرسیون لجستیک ترتیبی ۲۶۶
۴.۷. رگرسیون لجستیک چندجمله‌ای جریمه‌شده ۲۷۰
۱.۴.۷. مثال‌های گویا برای رگرسیون لجستیک چندجمله‌ای جریمه‌شده ۲۷۳
۵.۷. رگرسیون پواسون جریمه‌شده ۲۷۸
۶.۷. دیدگاه‌های نهایی ۲۸۳
منابع فصل هفتم ۲۹۸
فصل هشتم فضاهای هیلبرت هستهٔ بازتولیدکننده و روش‌های رگرسیون و طبقه‌بندی ۲۹۹
۱.۸. فضاهای هیلبرت هستهٔ بازتولیدکننده (RKHS) ۲۹۹
۲.۸. مدل هستهٔ تعمیم‌یافته ۳۰۱
۱.۲.۸. تخمین پارامتر تحت پارادایم مکرر ۳۰۲
۲.۲.۸. هسته‌ها ۳۰۴
۳.۲.۸. ترفند هسته ۳۰۶
۴.۲.۸. توابع محبوب هسته ۳۱۰
۵.۲.۸. فرایند دومرحله‌ای مجزا برای ساخت ماشین‌های هسته ۳۱۹
۳.۸. روش‌های هسته برای متغیرهای پاسخ گاوسی ۳۲۰
۴.۸. روش‌های هسته برای متغیرهای پاسخ باینری ۳۲۲
۵.۸. روش‌های هسته برای متغیرهای پاسخ طبقه‌ای ۳۲۳
۶.۸. مدل خطی مختلط با هسته ۳۲۴
۷.۸. تنظیم ابرپارامتر برای ساختن هسته‌ها ۳۲۸
۸.۸. روش‌های هستهٔ بیزی ۳۳۰
۱.۸.۸. پیش‌بینی توسعه‌یافتهٔ تحت هستهٔ بیزی BLUP ۳۳۳
۲.۸.۸. پیش‌بینی توسعه‌یافتهٔ تحت هستهٔ بیزی BLUP با یک متغیر پاسخ باینری ۳۳۵
۳.۸.۸. پیش‌بینی توسعه‌یافتهٔ تحت هستهٔ بیزی BLUP با یک متغیر پاسخ طبقه‌بندی‌شده ۳۳۷

۹.۸. هستهٔ بیزی چندویژگی ۳۳۸
۱۰.۸. روش‌های فشرده‌سازی هسته ۳۳۹
۱.۱۰.۸. پیش‌بینی‌کنندهٔ توسعه‌یافته تحت روش هستهٔ تقریبی ۳۴۵
۱۱.۸. دیدگاه‌های نهایی ۳۴۷
منابع فصل هشتم ۳۸۹
فصل نهم ماشین‌های بردار پشتیبان و رگرسیون بردار پشتیبان ۳۹۳
۱.۹. مقدمه‌ای بر ماشین بردار پشتیبان ۳۹۳
۲.۹. ابرصفحه ۳۹۵
۳.۹. طبقه‌بند حداکثر حاشیه ۳۹۷
۱.۳.۹. اشتقاق طبقه‌بندی‌کنندهٔ حداکثر حاشیه ۴۰۱
۲.۳.۹. ولف دوگانه ۴۰۴
۴.۹. مشتق از طبقه‌بندی بردار پشتیبان ۴۱۲
۵.۹. ماشین بردار پشتیبان ۴۱۷
۱.۵.۹. طبقه‌بندی یک در مقابل یک ۴۲۰
۲.۵.۹. طبقه‌بندی یک در مقابل همه ۴۲۱
۶.۹. رگرسیون بردار پشتیبان ۴۳۰
منابع فصل نهم ۴۴۰
فصل دهم مبانی شبکه‌های عصبی مصنوعی و یادگیری عمیق ۴۴۱
۱.۱۰. الهام گرفتن از مدل شبکهٔ عصبی ۴۴۱
۲.۱۰. اجزای ساختاری شبکه‌های عصبی مصنوعی ۴۴۴
۳.۱۰. توابع فعال‌سازی ۴۵۱
۱.۳.۱۰. خطی ۴۵۱
۲.۳.۱۰. واحد خطی اصلاح‌شده ۴۵۱
۳.۳.۱۰. Leaky ReLU ۴۵۳
۴.۳.۱۰. سیگموئید ۴۵۳
۵.۳.۱۰. Softmax ۴۵۴
۶.۳.۱۰. تانژانت هایپربولیک ۴۵۴
۴.۱۰. قضیهٔ تقریب عمومی ۴۵۵
۵.۱۰. توپولوژی‌های شبکهٔ عصبی مصنوعی ۴۵۷
چارچوب عصبی مصنوعی ۴۵۷
ساختار ارتباطی ۴۵۸
شبکهٔ پیش‌خور ۴۵۹
شبکه‌های بازگشتی ۴۵۹

۶.۱۰. کاربردهای موفق شبکهٔ عصبی مصنوعی و یادگیری عمیق ۴۶۰
۷.۱۰. توابع زیان ۴۶۳
۱.۷.۱۰. توابع زیان برای نتایج پیوسته ۴۶۴
۲.۷.۱۰. توابع زیان برای نتایج باینری و ترتیبی ۴۶۶
۳.۷.۱۰. توابع زیان منظم ۴۶۷
۴.۷.۱۰. روش توقف زودهنگام آموزش ۴۷۱
۸.۱۰. الگوریتم کینگ برای آموزش شبکه‌های عصبی مصنوعی: پس‌انتشار ۴۷۳
۱.۸.۱۰. الگوریتم پس‌انتشار: نسخهٔ آنلاین ۴۷۸
۱.۱.۸.۱۰. بخش پیش‌خور ۴۷۸
۲.۱.۸.۱۰. بخش پس‌انتشار ۴۷۸
۲.۸.۱۰. مثال گویای ۱.۱۰: با محاسبهٔ دستی ۴۷۹
۳.۸.۱۰. مثال گویای ۲.۱۰: با محاسبهٔ دستی ۴۸۳
منابع فصل دهم ۴۸۹
فصل یازدهم شبکه‌های عصبی مصنوعی و یادگیری عمیق برای پیش‌بینی ژنومی نتایج پیوسته ۴۹۱
۱.۱۱. ابرپارامترهایی که باید در شبکهٔ عصبی مصنوعی و یادگیری عمیق تنظیم شوند ۴۹۱
۱.۱.۱۱. توپولوژی شبکه ۴۹۱
۲.۱.۱۱. توابع فعال‌سازی ۴۹۲
۳.۱.۱۱. توابع زیان ۴۹۲
۴.۱.۱۱. تعداد لایه‌های پنهان ۴۹۳
۵.۱.۱۱. تعداد نورون‌ها در هر لایه ۴۹۴
۶.۱.۱۱. نوع منظم‌سازی ۴۹۶
۷.۱.۱۱. میزان یادگیری ۴۹۷
۸.۱.۱۱. تعداد دوره‌ها و دسته‌ها ۴۹۹
۹.۱.۱۱. طرح نرمال‌سازی برای داده‌های ورودی ۵۰۰
۲.۱۱. چارچوب‌های محبوب یادگیری عمیق ۵۰۱
۳.۱۱. بهینه‌سازها ۵۰۲
۴.۱۱. مثال‌های گویا ۵۰۵
منابع فصل یازدهم ۵۵۰
فصل دوازدهم شبکه‌های عصبی مصنوعی و یادگیری عمیق برای پیش‌بینی ژنومی نتایج باینری، ترتیبی و پیچیده ۵۵۱
۱.۱۲. آموزش شبکهٔ عصبی عمیق با نتایج باینری ۵۵۱
۲.۱۲. آموزش شبکهٔ عصبی عمیق با نتایج طبقه‌بندی‌شده (ترتیبی) ۵۵۷
۳.۱۲. آموزش شبکهٔ عصبی عمیق با شمارش نتایج ۵۶۲

۴.۱۲. آموزش شبکهٔ عصبی عمیق با نتایج چندمتغیره ۵۶۶
۱.۴.۱۲. شبکهٔ عصبی عمیق با نتایج پیوستهٔ چندمتغیره ۵۶۷
۲.۴.۱۲. شبکهٔ عصبی عمیق با نتایج باینری چندمتغیره ۵۷۱
۳.۴.۱۲. شبکهٔ عصبی عمیق با نتایج ترتیبی چندمتغیره ۵۷۵
۴.۴.۱۲. شبکهٔ عصبی عمیق با نتایج شمارشی چندمتغیره ۵۷۹
۵.۴.۱۲. شبکهٔ عصبی عمیق با نتایج ترکیبی چندمتغیره ۵۸۲
منابع فصل دوازدهم ۶۱۴
فصل سیزدهم شبکه‌های عصبی پیچشی ۶۱۷
۱.۱۳. اهمیت شبکه‌های عصبی پیچشی ۶۱۷
۲.۱۳. تنسورها ۶۱۹
۳.۱۳. پیچشی ۶۲۴
۴.۱۳. ادغام ۶۲۸
۵.۱۳. عملیات پیچشی برای تنسور یک‌بعدی در داده‌های ترتیبی ۶۳۲
۶.۱۳. انگیزهٔ استفاده از شبکه‌های عصبی پیچشی ۶۳۳
۷.۱۳. دلیل ترجیح شبکه‌های عصبی پیچشی، برای پردازش تصاویر، بر شبکه‌های عصبی عمیق پیش‌خور ۶۳۶
۸.۱۳. مثال‌های گویا ۶۴۲
۹.۱۳. مثال پیچش دوبعدی ۶۵۰
۱۰.۱۳. انتقاد از یادگیری عمیق ۶۵۸
منابع فصل سیزدهم ۶۶۹
فصل چهاردهم رگرسیون تابعی ۶۷۱
۱.۱۴. اصول تحلیل‌های رگرسیون خطی تابعی ۶۷۱
۲.۱۴. توابع پایه ۶۷۶
۱.۲.۱۴. پایهٔ فوریه ۶۷۶
۲.۲.۱۴. پایهٔ B-Spline ۶۷۷
۳.۱۴. مثال‌های گویا ۶۸۱
۴.۱۴. رگرسیون تابعی با تابع ضریب هموار ۶۹۱
۵.۱۴. برآورد بیزی از رگرسیون تابعی ۶۹۷
منابع فصل چهاردهم ۷۲۸
فصل پانزدهم جنگل تصادفی برای پیش‌بینی ژنومی ۷۲۹
۱.۱۵. انگیزهٔ استفاده از جنگل تصادفی ۷۲۹
۲.۱۵. درختان تصمیم ۷۳۰
۳.۱۵. جنگل تصادفی ۷۳۴

۴.۱۵. الگوریتم جنگل تصادفی برای متغیرهای پاسخ مستمر، باینری و طبقه‌ای ۷۳۶
۱.۴.۱۵. قوانین تقسیم ۷۳۹
۵.۱۵. الگوریتم جنگل تصادفی برای تعداد متغیرهای پاسخ ۷۴۹
۶.۱۵. الگوریتم جنگل تصادفی برای پاسخ‌های چندمتغیره ۷۵۴
۷.۱۵. دیدگاه‌های نهایی ۷۵۹
منابع فصل پانزدهم ۷۸۱

واژه‌نامهٔ فارسی ـ انگلیسی ۷۸۳
واژه‌نامهٔ انگلیسی ـ فارسی ۷۸۷
نمایه ۷۹۱

پیشگفتار مترجمان

یادگیری ماشین از زیرمجموعه‌های هوش مصنوعی است که به ماشین امکان می‌دهد به‌صورت خودکار از داده‌های گذشته، بدون برنامه‌نویسی صریح، یاد بگیرد. هوش مصنوعی گونه‌ای فنّاوری است که به ماشین امکان شبیه‌سازی رفتار انسان را می‌دهد. هدف هوش مصنوعی ایجاد سیستم رایانه‌ای هوشمند، مانند انسان، برای حل مشکلات پیچیده است. یادگیری ماشین الگوها را در داده‌ها پیدا می‌کند و آن‌ها را برای پیش‌بینی به کار می‌گیرد. با استفاده از نرم‌افزارهای معمول، مشخص می‌کنیم رایانه چه کاری انجام دهد و با استفاده از داده‌هایی که به آن وارد می‌شود، چگونه پاسخ را پردازش کند. حجم داده‌های در دسترس پی‌درپی رو به افزایش است. ماشین‌ها از این داده‌ها برای یادگیری و بهبود نتایج و ارائهٔ آن‌ها استفاده می‌کنند. این نتایج می‌تواند در بیان بینش‌های ارزشمند و همچنین اتخاذ تصمیمات تجاری آگاهانه بسیار مفید باشد. یادگیری ماشین روش منحصربه‌فرد مناسبی برای مقابله با چالش‌های حوزهٔ ژنتیک، اصلاح نژاد و بیوتکنولوژی گیاهی، دام و طیور است. این شیوه پیش‌بینی‌های دقیق، تصمیم‌گیری خودکار و ایجاد راه‌حل‌های نوآورانه برای مسائل پیچیده در علوم گیاهی و دامی را، با استفاده از الگوریتم‌هایی که می‌توانند الگوها را از داده‌ها بیاموزند، امکان‌پذیر می‌کند. برخلاف روش‌های آماری رایج، که اغلب خطی بودن و استقلال را در بین متغیرها فرض می‌کنند، یادگیری ماشین قادر است روابط و تعاملات غیرخطی بین عوامل ژنومی، محیطی و فنوتیپی را ثبت کند. یکی از مزیت‌های مهم این شیوه توانایی تحلیل داده‌های دارای ابعاد بالاست. در آیندهٔ نزدیک که اطلاعات ژنوتیپی گسترده یا حجم بسیار بالای اطلاعات توالی‌یابی ژنومی در دسترس قرار خواهد گرفت (و توانایی روش‌های رایج به چالش کشیده خواهد شد)، یادگیری ماشین به‌خوبی از عهدهٔ تحلیل چنین داده‌هایی بر خواهد آمد. استخراج روابط پیچیدهٔ بین متغیرها، مانند تأثیرات متقابل بین نشانگرها، نیز از دیگر مزیت‌های مطلوب این روش‌هاست. کاربرد این شیوه‌ها در پیش‌بینی ارزش‌های اصلاحی ژنومی به چند سال اخیر محدود می‌شود اما، به‌دلیل ویژگی‌های مطلوب و بهینه‌ای که دارند، استفاده از آن‌ها روزبه‌روز در حال گسترش است. روش‌های برآورد ارزش‌های اصلاحی در دو دستهٔ پارامتریک و ناپارامتریک قرار می‌گیرد. روش‌های پارامتریک شامل بهترین

پیش‌بینی ناریب خطی (BLUP)، GBLUP، روش‌های بیزی و دیگر موارد می‌شود و روش‌های ناپارامتریک نیز شبکه‌های عصبی، هوش مصنوعی و یادگیری ماشین را دربرمی‌گیرد. توسعهٔ الگوریتم‌های داده‌کاوی درزَمینهٔ اَبرداده‌ها در برآورد ارزش‌های اصلاحی نقش چشمگیری خواهد داشت. مجموعه‌ای از تکنولوژی‌ها همچون یادگیری ماشین، هوش مصنوعی و یادگیری عمیق (DL) در عصر جدید، در مقایسه با روش‌های قدیمی‌تر، برای بررسی صفات اقتصادی با ساختار پیچیده فرصت‌های مناسبی فراهم ساخته‌اند. یکی از تأثیرگذارترین کاربردهای یادگیری ماشین انتخاب ژنومی است که ارزش‌های اصلاحی را، براساس نشانگرهای ژنتیکی، پیش‌بینی می‌کند. روش‌های معمول مانند بهترین پیش‌بینی ناریب خطی با الگوریتم‌های یادگیری ماشین، ازجمله ماشین‌های بردار پشتیبان (SVM)، جنگل‌های تصادفی (RFs) و تکنیک‌های یادگیری عمیق، مانند شبکه‌های عصبی پیچشی، جایگزین می‌شوند. این الگوریتم‌ها دقت بالاتری دارند؛ به‌ویژه درمورد صفاتی که تحت تأثیر فعل‌وانفعالات پیچیدهٔ ژنتیکی و محیطی‌اند. یکی از مزایای این کتاب تا حدی مستقل بودن فصل‌های آن است. کتاب حاضر راهنمای مرجعی منحصربه‌فرد برای اشخاصی است که به روش‌های پیش‌بینی، به‌ویژه دربارهٔ گیاهان، علاقهٔ علمی یا حرفه‌ای دارند. امروزه شرکت‌های گوناگونی در سطح جهان از روش‌های یادگیری ماشین درزَمینهٔ مدیریت بهتر عملکردهایشان بهره می‌برند. این مسئله جریانی زودگذر نیست و کارآیی یادگیری ماشین کاملاً اثبات شده است. کتاب در دستْ دانش کافی درمورد الگوریتم‌های یادگیری تحت نظارت و بدون نظارت را در اختیار قرار می‌دهد. مزیت ویژهٔ آن روش ساده، خلاصه‌گویی و استفاده از مثال‌های عملی با استفاده از نرم‌افزار R برای درک بهتر دانشجویان است. نویسندگان این کتاب قصد داشته‌اند تکنیک‌ها و اطلاعات پایه‌ای یادگیری ماشین را در اختیار پژوهشگران، دانشجویان و کارشناسان قرار دهند. اصول پایه و مبانی اولیه‌ای در آن مطرح شده است که، درزَمینهٔ یادگیری ماشین، برای دانشجویان ضروری به نظر می‌رسد. کتاب شامل داده‌های واقعی گندم و ذرت، برگرفته از مرکز بین‌المللی تحقیقات ذرت و گندم (CIMMYT)، است. این داده‌ها در تعمیم روش‌های پیش‌بینی ژنومی به صفات و محیط‌های متعدد مؤثر بوده‌اند. دستاورد درخور توجه کتاب حاضر این است که نویسندگان توانسته‌اند طیف کاملی از روش‌های پیش‌بینی ژنومی را، درمورد داده‌های گیاهی، مطرح کنند. این کتاب نظریه و عملکرد پیش‌بینی‌های ژنومی را درمورد محیط‌های منفرد و چندگانه، صفات منفرد و چندگانه، برای داده‌های پیوسته و همچنین داده‌های باینری، طبقه‌بندی و شمارشی، مطرح می‌کند. درمورد «روش‌های آماری معمول»، مدل‌های مختلط در کنار رگرسیون‌های جریمه‌شده و روش‌های بیزی توضیح داده شده است. برای تکنیک‌های یادگیری ماشینی، روش‌های هسته و ماشین‌های بردار پشتیبانی و جنگل‌های

تصادفی را پیدا می‌کنیم. کاربردهای نظری این روش‌ها برای اصلاح‌کنندگان گیاهان نسبتاً پیشرفته است اما این کتاب درعین‌حال مطالب مقدماتی بسیاری را نیز درزَمینهٔ فلسفهٔ یادگیری آماری، مجموعه‌های آموزشی و آزمایشی، اعتبارسنجی متقابل، ماتریس‌ها، جریمه‌ها، اولویت‌ها، هسته‌ها، و شناسایی درختان، با دقت بیشتر و عمیق‌تر آموزش می‌دهد. یکی دیگر از نکات شایان اهمیت این کتاب گنجاندن برنامه‌های R برای تحلیل داده‌هاست. بسته‌های اصلی برای پیش‌بینی عبارت است از BGLR، BMTME، glmnet و کراس (keras) که دو مورد نخست کاربردی‌تر از دیگر بسته‌های مطرح‌شده‌اند. این کتاب فراتر از صرفاً بیان روش‌هایی درزَمینهٔ پیش‌بینی ژنومی است. بیشتر روش‌های پیش‌بینی بیان‌شده را می‌توان به‌راحتی درمورد دیگر انواع پیش‌بینی نیز به کار برد؛ بدین‌صورت صفات هدف ثابت باقی می‌ماند اما ورودی‌ها، به‌جای SNPها (چندشکلی تک‌نوکلئوتیدی)، شامل متابولیت‌ها، پروتئین‌ها، بیان ژن‌ها، متیلاسیون‌ها و دیگر داده‌های اومیکس است. همچنین می‌توان پیش‌بینی‌کننده‌های گوناگونی را با هم ترکیب کرد؛ مانند روش‌های هسته (kernel) که ورودی‌های متنوعی برای تعریف هسته‌های متعدد در آن استفاده می‌شود. این هسته‌ها به ماتریس‌های شباهت ژنوتیپی (تبدیل‌شده) بدل می‌شوند که ساختار تأثیرات تصادفی ژنوتیپی را تعیین می‌کنند. به‌طور مثال، ممکن است دو اثر تصادفی ژنوتیپی تعریف شود: یکی با استفاده از ماتریس روابط مبتنی‌بر SNP و دیگری با استفاده از ماتریس روابط مبتنی‌بر متابولیت‌ها ساختاردهی شود. مثال دیگر افزودن هسته‌ای غالب به مدلی ترکیبی است که، پیش‌تر از این، هسته‌ای برای تأثیرات افزایشی را شامل می‌شود. برای گنجاندن تأثیرات اپیستاتیک تمامی ردیف‌ها، علاوه‌بر تأثیرات افزایشی، می‌توان مدل‌های فضاهای هیلبرت هستهٔ بازتولیدکننده را به کار برد. برای استفاده از اطلاعات فنوتیپی با توان عملیاتی بالا به‌منزلهٔ پیش‌بینی‌کنندهٔ صفات فنوتیپی، روش‌های رگرسیون تابعی شرح داده شده است و در نهایت، زمینهٔ استفادهٔ آسان از تصاویر، برای پیش‌بینی ویژگی‌های گیاهی ازسوی زبان‌آموزان، مهیا شده است.

عباس سعیدی، عضو هیئت علمی دانشگاه شهید بهشتی

زهره حاجی‌برات

زهرا حاجی‌برات

پاییز ۱۴۰۵

سخن آغازین

درزَمینهٔ اصلاح نباتات، مفهوم پیش‌بینی همیشه نقشی اساسی داشته است. مدت‌های طولانی، اصلاح‌کنندگان سعی می‌کردند عملکرد ژنتیکی برتر، ارزش ژنوتیپی یا ارزش اصلاحی را براساس سیستم تولیدمثل گیاه، طبق مشاهدات روی فنوتیپ مورد نظر، با پیش‌بینی‌های مرتبط با نسبت واریانس ژنتیکی به فنوتیپی حدس بزنند. این نسبت معمولاً وراثت‌پذیری براساس میانگین خطی بود. حدود سال ۱۹۹۰ که نشانگرهای مولکولی در دسترس قرار گرفت، تقریباً بلافاصله پیش‌بینی مقادیر ژنوتیپی طبق مدل‌های خطی مختلط حاوی ماتریس‌های طراحی یا پیش‌بینی‌کنندهٔ[1] ساخته‌شده روی اطلاعات نشانگر ایجاد شد. مقالهٔ رکس برناردو[2] در سال ۱۹۹۴ (*Crop Science*, 34) احتمالاً اولین مقاله‌ای بود که این موضوع را درمورد گیاهان پیشنهاد کرد. ایدهٔ مشابهی با نام پیش‌بینی و انتخاب ژنومی (GS)[3]، پس‌از انتشار مقالهٔ مشهور تئو میویسن[4]، بن هایز[5] و مایک گدارد[6] در شمارهٔ ۱۵۷ مجلهٔ *ژنتیک*[7] (۲۰۰۱)، در اصلاح نژاد حیوانات انقلابی ایجاد کرد.

در مقایسه با موفقیت سریع در اصلاح جانوران، معرفی و پذیرش پیش‌بینی ژنومی در اصلاح گیاهان مدت بیشتری طول کشید. ممکن است دلیل این تفاوت بین اصلاح نژاد جانوران و گیاهان آن باشد که برخلاف اصلاح نژاد حیوانات، در اصلاح نباتات، شناسایی جایگاه‌های کمّی مفید بسیاری از صفات (QTLs) چندان دشوار نبود. بنابراین، در اصلاح نباتات با دست‌کاری آلل‌های QTL در روش‌های انتخاب به‌کمک نشانگر، پیشرفتی ممکن شد. دلیل دیگر اجرای کندتر پیش‌بینی ژنومی، در گیاهان، مرتبهٔ کوچک‌تر قدر مجموعه‌های آموزشی است. در اصلاح حیوانات، ۱۰۰۰ یا ۱۰۰۰۰ عددی رایج است اما، درمورد گیاهان، ۱۰۰۰ نقطه‌ای پرت در انتهای کار است.

1. predictor
2. Rex Bernardo
3. genomic selection
4. Theo Meuwissen
5. Ben Hayes
6. Mike Goddard
7. *Genetics*

عارضهٔ مهم دیگر، در روش‌های پیش‌بینی ژنومی برای گیاهان، فراگیر بودن ژنوتیپ برآثر برهم‌کنش محیطی است.

طی بیست سال گذشته، روش‌های بسیاری برای پیش‌بینی ژنومی مطرح شده است. ازآنجاکه هدف این روش‌ها استفاده از چندشکلی‌های تک‌نوکلئوتیدی (SNP)[1] (یا سایر نشانگرها/ اطلاعات توالی) برای پیش‌بینی مقادیر ژنوتیپی است، تعداد پیش‌بینی‌کننده‌ها معمولاً بیشتر و اغلب بسیار بیشتر از تعداد ژنوتیپ‌های دارای فنوتیپ، یعنی اندازه، است. بنابراین برخی شیوه‌های منظم‌سازی[2] مجموعهٔ آموزشی در روش برآورد تأثیرات SNP ضروری است. این منظم‌سازی ممکن است ازجهاتی برای اصلاح‌کنندگان گیاه آشنا باشد. مدل خطی مختلط کلاسیک برای پیش‌بینی ارزش‌های ژنوتیپی از میانگین خطی وراثت‌پذیری و تأثیرات ژنوتیپی تصادفی، که براساس روابط شجره‌نامه ساختار می‌شوند، بهره می‌برد و می‌توان آن را به خطی مشابه و آمیخته با آثار ژنوتیپی، که با روابط نشانگر بین ژنوتیپ‌ها ساختار یافته است، تعمیم داد. در آن صورت، حرکت ما از بهترین پیش‌بینی‌های نااریب خطی کلاسیک یا BLUP به‌سمت BLUPهای فعال‌شده با ژنوم یا G-BLUP خواهد بود.

G-BLUP روش محوری پیش‌بینی ژنومی در اصلاح حیوانات و گیاهان به شمار می‌رود. به‌دلایل محاسباتی، هر زمان که تعداد ژنوتیپ‌ها کمتر از تعداد نشانگرها باشد، فرمول مدلی ترکیبی که روابط بین ژنوتیپ‌ها را براساس پروفایل‌های نشانگر شکل می‌دهد مطلوب‌تر از فرمولاسیونی است که ماتریس طراحی حاوی SNPها را درج می‌کند. فرمول اخیر شکلی از رگرسیون ریج است و پیش‌بینی‌های آن RR-BLUP نامیده می‌شود. G-BLUP و RR-BLUP را می‌توان، براساس دیدگاه انقباض، در مدلی مختلط توضیح داد اما، به‌صورت جایگزین، در نظر گرفتن انگیزه‌هایی از رگرسیون جریمه‌ای یا رویکرد بیزی نیز درمورد آن‌ها امکان‌پذیر است. RR-BLUP و G-BLUP از جریمهٔ ریج[3] استفاده می‌کنند و درعین‌حال می‌توان آن‌ها را براساس استفاده از پیشینی عادی روی تأثیرات SNP نیز شناخت. در این مرحله، پیشنهاد انواع جریمه‌ها و مقدمات دیگری در پیش‌بینی ژنومی، با ایدهٔ درج اطلاعات درمورد ساختار ژنتیکی صفت، بدیهی به نظر می‌رسد. برای نمونه، جریمه‌ها و اولویت‌ها می‌توانند میزانی از پراکندگی، مانند جریمه‌های

1. single nucleotide polymorphism
2. regularization
3. Ridge penalty

نوع لاسو[1] و اولویت‌های نمایی لاپلاس/ دوبل[2]، را القا کنند. علاوه‌بر این، درمورد توزیع تأثیرات SNP (غیرصفر)، مفروضات گوناگونی را می‌توان ایجاد کرد. آثار SNP را می‌توان در جهتی دیگر، به توزیع‌های گوناگون با واریانس‌های متفاوت، مانند بزرگی‌های متفاوت تأثیرات SNP، اختصاص داد.

کیت ابزار برای پیش‌بینی مقادیر ژنوتیپی، یا بهتر از آن، عملکرد فنوتیپی مورد انتظار از نمایه‌های نشانگر، به‌سرعت با مدل‌های مخلوط خطی (G-BLUP، RR-BLUP)، رگرسیون‌های جریمه‌شده (ریج، لاسو)، و روش‌های بیزی (A، B، Cπ، R) مجهز شد. د لُس کامپوس[3] و همکاران (۲۰۱۳)، در شمارهٔ ۱۹۳ *ژنتیک*، تمامی این روش‌ها را به‌طور کامل بررسی کردند. علاوه‌بر روش‌های پیش‌بینی «آماری» فوق، انواع تکنیک‌های یادگیری ماشین مانند ماشین‌های بردار پشتیبان، جنگل‌های تصادفی و یادگیرنده‌های عمیق برای پیش‌بینی ژنومی پیشنهاد شده‌اند. عضدی[4] و همکاران (*G3*, 2019) روش‌های آماری و یادگیری ماشینی را کاملاً بررسی و مقایسه کردند. نتیجهٔ این مقاله و بسیاری مقالات دیگر، درمورد مقایسهٔ روش‌های پیش‌بینی ژنومی، تا حدودی اطمینان‌بخش و به همان میزان ناامیدکننده است و بیان می‌کند دقت پیش‌بینی بین روش‌ها تغییر چندانی ندارد.

نویسندگان کتاب پیش رو، در ارزیابی طیف وسیعی از روش‌های پیش‌بینی در گیاهان براساس G-BLUP تا یادگیری عمیق[5]، نقش مهمی داشته‌اند. تحقیقات آن‌ها داده‌های واقعی گندم و ذرت از مؤسسهٔ بین‌المللی مشهور اصلاح نژاد با نام CIMMYT، مرکز بین‌المللی بهبود ذرت و گندم، را دربرمی‌گیرد. این نویسندگان در تعمیم روش‌های پیش‌بینی ژنومی به صفات و محیط‌های متعدد نیز مؤثر بودند. مسیر انتخابی در این کتاب برمبنای تعمیم مدل‌های مختلط چندویژگی[6] و چندمحیطی، با تعریف ساختاری برای ژنوتیپ تصادفی × تأثیرات صفت و ژنوتیپ × تأثیرات محیطی به‌صورت محصول کرونکر[7] از ساختار ماتریس‌ها در ژنوتیپ‌ها، ازیک‌سو، صفات یا محیط‌ها و ازسوی‌دیگر، محیط‌ها بود. برای ژنوتیپ‌ها، روابط با استفاده از نشانگرها تعریف شد. برای صفات یا محیط‌ها، می‌توان ماتریس‌های تحلیلی واریانس- کوواریانس بدون ساختار یا عاملی را انتخاب

1. Lasso
2. Laplace/ double
3. de los Campos
4. Azodi
5. deep learning
6. multi-trait
7. Kronecker

کرد. طبق توضیحات یارکین[1] و همکاران در شمارهٔ ۱۲۷ مجلهٔ *ژنتیک نظری و کاربردی*[2] (۲۰۱۴)، هنگامی که مشخصات محیطی صریح در دسترس باشد، روابط بین محیط‌ها نیز می‌تواند براساس شباهت‌های محیطی باشد. یکی دیگر از پژوهش‌های مشارکتی مهم نویسندگان کتاب بیان نحوهٔ پیش‌بینی ژنومی صفات باینری، صفات طبقه‌بندی و داده‌های شمارش بوده است. راه پیش رو، درمورد چنین داده‌هایی، اجرای مدل‌های خطی تعمیم‌یافتهٔ بیزی (ترکیبی) است.

دستاورد شایان توجه کتاب حاضر این است که نویسندگان توانسته‌اند طیف کاملی از روش‌های پیش‌بینی ژنومی را درمورد داده‌های گیاهی ارائه دهند. این کتاب نظریه و عمل پیش‌بینی‌های ژنومی را برای محیط‌های منفرد و چندگانه، صفات منفرد و چندگانه، درمورد داده‌های پیوسته و همچنین داده‌های باینری، طبقه‌بندی و شمارشی ارائه می‌دهد. دربارهٔ «روش‌های آماری رایج»، مدل‌های مختلط در کنار رگرسیون‌های جریمه‌شده و روش‌های بیزی توضیح داده شده‌اند. برای تکنیک‌های یادگیری ماشین، روش‌های هسته، ماشین‌های بردار پشتیبانی، جنگل‌های تصادفی و انواع یادگیرندگان قدیمی و مدرن ماشینی و عمیق را پیدا می‌کنیم و در این بین، با رگرسیون تابعی مواجه می‌شویم. شیوه‌های نظری پرداختن به روش‌ها برای اصلاح‌کنندگان متوسط گیاهان نسبتاً پیشرفته است اما این کتاب، هم‌زمان، مطالب مقدماتی بسیاری را دربارهٔ فلسفهٔ نهفته در پسِ یادگیری آماری، مبادلات بایاس واریانس، مجموعه‌های آموزشی و آزمایشی، اعتبارسنجی متقابل (CV)[3]، ماتریس‌ها، جریمه‌ها، پیشین‌ها، هسته‌ها، درخت‌ها و یادگیرندگان عمیق مطرح می‌کند؛ این به خوانندگانی که آموزش آماری کمتری دیده‌اند امکان می‌دهد بیشتر تکنیک‌ها را تا حدی درک کنند. یکی دیگر از نکات مثبت استفاده از کتاب گنجاندن برنامه‌های R فراوان برای تحلیل خودتان است. بسته‌های[4] مرکزی برای پیش‌بینی عبارت‌اند از BGLR، BMTME، glmnet و کراس[5] که دو مورد اول بیشتر در حیطهٔ تخصص نویسندگان این کتاب است. احتمال دارد دغدغه‌هایی درمورد امکان‌سنجی محاسباتی اجراهای بیزی برای مجموعه داده‌های بزرگ‌تر، یعنی صد ژنوتیپ و سی صفت/ آزمایش یا بیشتر، نیز مطرح شود.

در این کتاب، مطالبی فراتر از روش‌های پیش‌بینی ژنومی آمده است. بیشتر روش‌های پیش‌بینی را می‌توان به‌راحتی با انواع دیگر پیش‌بینی‌ها تطبیق داد. در این پیش‌بینی‌ها، صفات هدف ثابت

1. Jarquín
2. *Theoretical and Applied Genetics*
3. cross-validation
4. package
5. Keras

می‌ماند اما ورودی‌ها SNP نیست بلکه متابولیت‌ها، پروتئین‌ها، بیان ژن‌ها، متیلاسیون‌ها و داده‌های دیگر از نوع اومیکس[1] است. مانند روش‌های هسته، که ورودی‌های گوناگونی برای تعریف هسته‌های متعدد به کار می‌رود و به ژنوتیپ (تبدیل‌شده) × ماتریس‌های تشابه ژنوتیپ ترجمه می‌شوند که تأثیرات تصادفی ژنوتیپی را ساختار می‌دهند، پیش‌بینی‌کننده‌های گوناگون نیز می‌توانند ترکیب شوند. برای نمونه، ممکن است دو اثر ژنوتیپی تصادفی تعریف شود؛ با ماتریس رابطه‌ای مبتنی‌بر SNP که اولین اثر ژنوتیپی تصادفی را شکل می‌دهد و ماتریس رابطه‌ای مبتنی‌بر متابولیت که اثر ژنوتیپی دوم را ساختار می‌دهد. مثال دیگر شامل افزودن هسته‌ای غالب به مدلی ترکیبی است که، از پیش، هسته‌ای برای تأثیرات افزایشی را دربرمی‌گیرد. برای گنجاندن آثار اپیستاتیک[2] تمامی ردیف‌ها، علاوه‌بر تأثیرات افزایشی، می‌توان از مدل‌های فضای هیلبرت هستهٔ بازتولیدکننده[3] استفاده کرد. برای استفاده از اطلاعات فنوتیپی با توان عملیاتی بالا به‌منزلهٔ پیش‌بینی‌کنندهٔ صفات فنوتیپی، روش‌های رگرسیون تابعی توضیح داده شده‌اند. استفاده از تصاویر سبب تسهیل پیش‌بینی صفات گیاهی ازسوی یادگیرنده‌های عمیق می‌شود.

رویکرد این کتاب ارائهٔ دایرةالمعارفی درزَمینهٔ پیش‌بینی اصلاح نباتات است. درمورد هر روش امکان‌پذیر با ویژگی‌های آماری، مثال‌ها و نرم‌افزار به‌صورت مشروح توضیح داده و تقریباً بیشتر برنامه‌های رایج پوشش داده شده است. فصل‌های کتاب تقریباً مستقل‌اند و این نکته سبب افزایش کاربردی بودن و سودمندی کتاب می‌شود. اثر پیش رو راهنمای مرجع منحصربه‌فردی برای افرادی است که علاقهٔ علمی یا حرفه‌ای به روش‌های پیش‌بینی درزَمینهٔ گیاهان دارند.

واخنینگن، هلند
فرد فن اِیوویک
۱ سپتامبر ۲۰۲۱

1. omics
2. Epistatic effects
3. Reproducing Kernel Hilbert Space

پیشگفتار نویسندگان

درزَمینهٔ اصلاح گیاهان، مدل‌های پیش‌بینی برای اجرای موفقیت‌آمیز انتخاب ژنومی (GS) ضروری‌اند. این روش از شیوه‌های پیش‌بینی در نظر گرفته می‌شود که برای آموزش مدل‌ها با استفاده از جمعیت آموزشی مرجعی، حاوی داده‌های فنوتیپی و ژنوتیپی شناخته‌شده، به کار می‌رود و سپس پیش‌بینی‌ها درمورد مجموعه داده‌ای آزمایشی، که فقط دربردارندهٔ داده‌های ژنومی است، انجام می‌شود. بااین‌حال، ازآنجاکه مدلی عمومی وجود ندارد (قضیهٔ ناهار رایگان)[1]، بسیاری از مدل‌ها باید برای مجموعه داده‌ای خاص ارزیابی شوند و بهترین گزینه برای هر موقعیت خاص انتخاب شود. بنابراین مدل‌های آماری بسیار متنوعی برای پیش‌بینی در انتخاب ژنومی به کار می‌رود. مدل‌های پیش‌بینی ماشین آماری مبتنی‌بر ژنوم GS با پیچیدگی‌های متعددی مواجه‌اند زیرا انتخاب مدل به عوامل متعددی، مانند معماری ژنتیکی صفت، تراکم نشانگر، اندازهٔ نمونه، گسترهٔ عدم تعادل لینکاژی و برهم‌کنش محیط ژنوتیپ، بازمی‌گردد.

درزَمینهٔ علوم گیاهی، محبوب‌ترین مدل یادگیری آماری مدل خطی مختلط است که از معادلات هندرسون[2] برای یافتن بهترین تخمین‌های نااریب خطی (BLUEs) درمورد تأثیرات ثابت و بهترین پیش‌بینی نااریب خطی (BLUPs) درزَمینهٔ تأثیرات تصادفی استفاده می‌کند و همچنین همتای بیزی این مدل که نویسندگان نسخه‌های گوناگونی را براساس آن توسعه دادند و به‌اصطلاح، روش‌های الفبای بیزی[3] [رگرسیون ریج بیزی (BRR)[4]، BayesA، BayesB، BayesC، Bayes Lasso (BL) و دیگر موارد] را ابداع کردند. چندین روش یادگیری ماشینی در دسترس است؛ ازجمله یادگیری عمیق، جنگل تصادفی و ماشین‌های بردار پشتیبانی. اجرای

1. free lunch theorem
2. Henderson
3. Bayesian Alphabet Methods
4. Bayesian Ridge Regression

ماشین‌های جنگل تصادفی و بردار پشتیبان بسیار آسان است زیرا به پارامترهای اندکی برای تنظیم نیاز دارند و به‌سرعت، پیش‌بینی‌های بسیار رقابتی را در اختیار می‌گذارند.

روش‌های ناشی از زمینه‌های آماری و یادگیری ماشینی را روش‌های یادگیری ماشین آماری می‌نامند.

تمامی روش‌های آماری یادگیری ماشین، برای ایجاد پیش‌بینی‌های معقول، به ورودی‌های بسیار پیش‌پردازش‌شده (مهندسی ویژگی) نیاز دارند. به‌عبارت‌دیگر، بیشتر روش‌های آماری یادگیری ماشینی به مداخلهٔ بیشتر کاربر در پیش‌پردازش ورودی‌ها نیاز دارند که باید به‌صورت دستی انجام شود. در فصل یکم، استفاده از داده‌های بزرگ برای پیش‌بینی و تخمین ازطریق یادگیری ماشین آماری و کاربردهای کلی آن در کشاورزی و ژنتیک ارزیابی می‌شود و عناصر اصلی انتخاب ژنومی و پتانسیل آن، برای بهبود گیاه، بیان می‌شود. فصل دوم به جزئیات آماده‌سازی داده‌های لازم برای اجرای روش‌های آماری یادگیری ماشین، به‌منظور انتخاب ژنومی، می‌پردازد. در این فصل، ابزارهایی برای تمیز کردن، تعیین و تشخیص محاسبات فراوانی آلل فرعی و اصلی، رمزگذاری مجدد نشانگر، فراوانی نشانگر ناهمگن، فراوانی NA و روش‌هایی برای محاسبهٔ ماتریس رابطهٔ ژنومی (GRM)[1] مطرح می‌شود. در فصل سوم، جزئیات مدل رگرسیون خطی چندگانه (LMRM)[2] با روش‌های کاهش گرادیان (GD)[3] برای یادگیری پارامترهای تحت این مدل می‌آید. مدل رگرسیون خطی چندگانه درمورد جریمه‌های ریج و لاسو مشتق شده است.

در هر فصل از این کتاب چندین مثال، ازجمله کدهای R لازم، آمده است که درک و استفادهٔ سریع آن را آسان می‌کند. فصل چهارم پدیدهٔ اضافهٔ برازش را توصیف می‌کند؛ این پدیده زمانی رخ می‌دهد که یک مدل یادگیری ماشین آماری خطا و همچنین سیگنال موجود در داده‌های آموزشی را به‌طور کامل یاد می‌گیرد. همچنین زمانی که فقط چند پیش‌بینی‌کننده در مدل یادگیری ماشین آماری گنجانده‌شده به مدل‌های دچار برازش کمتر اشاره می‌کند و این مدل نمایندهٔ چندان قدرتمندی برای ساختار کامل الگوی داده‌ها محسوب نمی‌شود. در فصل پنجم، چارچوب مدل خطی مختلط، با بررسی سه روش تخمین پارامتر (حداکثر درست‌نمایی، الگوریتم EM و REML)، توضیح داده و شیوهٔ پیش‌بینی‌های فعال‌شده با ژنومی براساس این چارچوب

1. genomic relationship matrix
2. linear multiple regression model
3. gradient descent

بیان می‌شود. در این فصل، کاربرد مدل‌های خطی مختلط درمورد مدل‌های خطی مختلط تک و چندویژگی نشان داده شده و کدهای R برای تحلیل‌ها بیان شده است.

فصل ششم پارادایم بیزی را برای تخمین پارامتر چندین رگرسیون خطی بیزی توصیف می‌کند که در پیش‌بینی فعال ژنومی استفاده می‌شود و در نمونه‌های متفاوت پیش‌بینی‌کننده‌های اجراشده در کتابخانهٔ رگرسیون خطی تعمیم‌یافتهٔ بیزی (BGLR)، درمورد متغیرهای پاسخ پیوسته، وقتی شامل تأثیرات اصلی (محیط‌ها و ژنوتیپ‌ها) می‌شوند، و همچنین اصطلاحات برهم‌کنش مرتبط با برهم‌کنش ژنوتیپ با محیط، وجود دارد. پارادایم‌های پیش‌بینی کلاسیک و بیزی برای داده‌های دسته‌بندی و شمارشی در فصل هفتم ترسیم شده‌اند. در این فصل، در کتابخانهٔ BGLR و glmnet که رگرسیون لجستیک چندجمله‌ای جریمه‌شده و رگرسیون پواسون جریمه‌شده ارائه شده‌اند، نمونه‌هایی اجرا شد.

مدل یادگیری ماشین نیمه‌پارامتریک شایان توجهی در فصل هشتم معرفی شده که برخی توابع از دست دادن براساس یک چارچوب مدل ثابت، با مثال‌هایی از متغیرهای پاسخ گاوسی، باینری و طبقه‌ای، را نشان می‌دهد. در این فصل، کاربرد مدل‌های مختلط با هسته‌ها بیان شده که شامل مثال‌هایی درمورد روش‌های رگرسیون چندویژگی RKHS برای متغیرهای پاسخ پیوسته است.

به منشأ و محبوبیت روش‌های ماشین بردار پشتیبان (SVM)[1]، در فصل نهم، اشاره و مشتقات آن‌ها، با استفاده از طبقه‌بندی‌کنندهٔ با حداکثر حاشیه[2] و طبقه‌بندی‌کنندهٔ بردار پشتیبان[3]، برجسته می‌شود؛ همچنین اصول اولیهٔ ساخت روش‌های هسته‌ای گوناگون، که در ماشین بردار پشتیبان مجازند، توضیح داده می‌شود. در فصل دهم، به پایه‌های شبکه‌های عصبی مصنوعی (ANN)[4] و روش‌های یادگیری عمیق و نیز موارد تأثیرگذاری اصلی در شیوهٔ ایجاد این روش‌ها پرداخته می‌شود. در این فصل، تابع فعال‌سازی را تعریف و به نقش آن در شناسایی الگوهای غیرخطی در داده‌های ورودی اشاره می‌کنیم. فصل یازدهم به عناصری به‌منظور اجرای شبکه‌های عصبی عمیق (یادگیری عمیق) (DNN)[5]، برای نتایج مستمر، می‌پردازد. چندین مثال عملی، با داده‌های اصلاح نباتات، برای اجرای شبکه‌های عصبی عمیق در کتابخانهٔ کراس بیان شده است. کارآیی شبکه‌های عصبی مصنوعی

1. support vector machine
2. maximum margin classifier
3. support vector classifier
4. artifitial neural network
5. deep neural network

و روش‌های یادگیری عمیق به پیش‌بینی مبتنی‌بر ژنوم در کراس، درمورد نتایج باینری، طبقه‌ای و مختلط براساس شبکه‌های پیش‌خور[1] (فصل دوازدهم)، گسترش یافته است. یادگیری عمیق اغلب در پیش‌بینی مبتنی‌بر ژنوم استفاده نمی‌شود زیرا برتری آن در پیش‌بینی عملکرد افراد مشاهده‌نشده، در قیاس با روش‌های معمولی یادگیری ماشین آماری، هنوز نامشخص است. باوجوداین کاربردهای یادگیری عمیق اندکی در پیش‌بینی‌های مبتنی‌بر ژنوم نشان می‌دهد که این روش انتخاب ژنوتیپ‌های کاندیدا را، در مراحل اولیهٔ برنامه‌های اصلاح، بهبود می‌بخشد و همچنین درک فرایندهای بیولوژیکی پیچیده در رابطهٔ بین فنوتیپ‌ها و ژنوتیپ‌ها را افزایش می‌دهد.

چندین مثال از آزمایش‌های اصلاح نباتات، ازجمله داده‌های ژنومی، بیان شده است. فصل سیزدهم انگیزه‌ای رسمی برای شبکه‌های عصبی پیچشی (CNN یا ConvNetsها)[2] مطرح می‌کند که مزایای این توپولوژی را، در مقایسه با شبکه‌های پیش‌خور برای پردازش تصاویر، نشان می‌دهد. چندین مثال عملی با داده‌های اصلاح نباتات با استفاده از CNN در دو سناریو ارائه شده است: الف) داده‌های ورودی یک‌بعدی و ب) داده‌های ورودی دوبعدی. مثال‌هایی نیز آمده است که نحوهٔ تنظیم ابرپارامترها در CNN را نشان می‌دهد.

روش‌های فنومیکس پیشرفته قادرند، با استفاده از دوربین‌های ابرطیفی، صدها نقطهٔ دادهٔ بازتابی را در نوارهای باریک و در محیط‌ها و مراحل متفاوت رشد گیاهان ارائه دهند. فنّاوری تعیین فنوتیپ اکنون می‌تواند، به‌سرعت و بادقت، داده‌هایی درمورد صفات زراعی را، براساس سرمایه‌گذاری‌ها و پیشرفت‌های گسترده در فنّاوری‌های تعیین فنوتیپ گیاهان، به دست آورد. هدف اصلی تعیین فنوتیپ با توان بالا کاهش هزینهٔ داده‌ها در هر قطعه و افزایش دقت پیش‌بینی‌های ژنومیکس در اوایل فصل رشد گیاهان است. فصل چهاردهم اصول نظری و مسائل عملی استفاده از رگرسیون تابعی درزَمینهٔ پیش‌بینی‌های فنومیکس (تصاویر) و ژنومیکس (نشانگرهای مولکولی متراکم) را مطرح می‌کند و رگرسیون تابعی داده‌ها را، ازطریق توابع پایه، نمایش می‌دهد. همچنین نسخهٔ بیزی رگرسیون تابعی استخراج و جزئیات اجرای آن توضیح داده می‌شود.

در فصل پانزدهم، توضیحاتی دربارهٔ الگوریتم جنگل تصادفی، ابرپارامترهای اصلی که نیاز به تنظیم دارند و همچنین قوانین تقسیم‌بندی گوناگون اصلی در اجرای آن آمده و درزَمینهٔ آموزش مدل‌های جنگل تصادفی با انواع متغیرهای پاسخ گوناگون، با استفاده از اصلاح گیاه و پیش‌بینی

1. feed-forward network
2. convolutional neural networks

مبتنی‌بر ژنوم، مثال‌هایی مطرح شده است. همچنین الگوریتم جنگل تصادفی برای نتایج چندمتغیره مطرح و مطلوب‌ترین قوانین تقسیم‌بندی آن توضیح داده می‌شود.

کولیما، مکزیک[1]
اُسوال آنتونیو مونته‌سینوس لوپز[2]
گوئادالاخارا[3]، مکزیک
آبلاردو مونته‌سینوس لوپز
تِسکوکو، ایالت مخیکو[4]، مکزیک
خوزه کروسا[5]

1. Colima, México
2. Osval Antonio Montesinos López
3. Guadalajara
4. Texcoco, Estado de México
5. José Crossa

فصل اول
عناصر رایج انتخاب ژنومی و یادگیری آماری

۱.۱. داده‌ها به‌منزلهٔ سلاحی قدرتمند

به‌لطف پیشرفت فنّاوری‌های دیجیتال، مانند دستگاه‌های الکترونیکی و شبکه‌ها، امکان دیجیتالی کردن بسیاری از مشاغل، فرایندها و خدمات ایجاد شده است که حجم عظیمی از داده‌ها را پدید می‌آورند. این «داده‌های بزرگ» منتقل، جمع‌آوری و تحلیل می‌شوند تا بینش عمیقی درمورد فرایندها و رفتار انسان حاصل شود. به همین دلیل، داده‌ها به «نفت جدید» مشهور شده‌اند زیرا «ارزش داده‌ها برای این قرن همانند نفت برای قرن گذشته است» ــ یعنی محرکی برای تغییر، رشد و موفقیت. همچنان که الگوریتم‌های آماری و یادگیری ماشینی اطلاعات را از داده‌های خام استخراج می‌کنند، اطلاعات می‌تواند برای ایجاد دانش استفاده شود و دانش به درک منجر می‌شود (Sejnowski 2018). ما ابزار و تخصص جمع‌آوری داده‌ها را، از منابع متفاوت و در هر قالبی، داریم که سنگ‌بنای استراتژی دادهٔ مدرن است و قدرت هوش مصنوعی[1] را نشان می‌دهد و هر روز حدود $۲٫۵\times۱۰^{۱۸}$ بایت داده ایجاد می‌کنیم (McKinsey Global Institute 2016). این بدان‌معنی است که تقریباً ۹۰٪ داده‌های جهان طی دو سال گذشته ایجاد شده است. این ظرفیت بی‌سابقهٔ تولید داده، ازطریق منابع متعددی مانند توئیت‌ها، یوتیوب، وبلاگ‌ها، حسگرها، اینترنت، گوگل، ایمیل‌ها، تصاویر، سبب افزایش ارتباطات و جریان داده‌های جهانی شده است. برای نمونه، گوگل بیش از چهل‌هزار جست‌وجو در هر ثانیه (و ۳٫۵میلیارد جست‌وجو در روز) را پردازش می‌کند، ۴۵۶هزار توئیت ارسال می‌شود، ۴۱۴۶۶۰۰ ویدئوی یوتیوب در دقیقه نمایش داده می‌شود، در هر دقیقه ۱۵۴۲۰۰ تماس اسکایپ برقرار می‌شود، ۱۵۶میلیون ایمیل فرستاده می‌شود و

1. artificial intelligence