Review: Stat/ML in a Nutshell

Advanced Topics in Statistical Learning, Spring 2024

Ryan Tibshirani

We will take a brief tour of core concepts in statistics and machine learning. Most of this will/should be familiar to you already. Refer to, e.g., Wasserman (2004); Hastie et al. (2009) for (much) more details.

1 Probability: modes of convergence and limit laws

Let X_n, $n = 1, 2, 3, \ldots$ be a sequence of real-valued random variables.

- Almost sure convergence: $X_n \xrightarrow{\text{as}} X$ means that $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$.
- Convergence in probability: $X_n \xrightarrow{p} X$ means that, for each $\epsilon > 0$, $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.
- Convergence in distribution: $X_n \xrightarrow{d} X$ means that $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$ at all continuity points x of the law of X.
- Almost sure convergence implies convergence in probability.
- Convergence in probability implies convergence in distribution.
- Convergence in distribution does *not* imply convergence in probability (except in the special case that the limiting distribution is a constant).
- Convergence in distribution is equivalent to $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded, continuous f (this, together with a collection of other equivalences, is often called the portmanteau lemma).
- Asymptotic probability notation: $X_n = O_p(a_n)$ means that X_n/a_n is bounded in probability, i.e., for each $\epsilon > 0$, there exists an $M > 0$ such that for sufficiently large n,

$$\mathbb{P}\left(\left|\frac{X_n}{a_n}\right| > M\right) \leq \epsilon.$$

  Meanwhile, $X_n = o_p(a_n)$ means that $X_n/a_n \xrightarrow{p} 0$.
- Law of large numbers (LLN): if $X_1, X_2, \ldots$ are i.i.d. with $\mu = \mathbb{E}[X_i]$, then for the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_n,$$

  it holds that $\bar{X}_n \xrightarrow{\text{as}} \mu$.

  Note: this implies $\bar{X}_n \xrightarrow{p} \mu$, but there are some cases where latter holds (convergence in probability) but the former does not (almost sure convergence), i.e., when the mean is not defined (infinite).
- Central limit theorem (CLT): under the same conditions, if additionally a second moment exists and we let $\sigma^2 = \text{Var}[X_i]$, then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1).$$

- Gilvenko-Cantelli theorem: if $X_1, X_2, \ldots$ are i.i.d. then for the empirical distribution function F_n defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n 1\{X_n \leq x\},$$

it holds that

$$\sup_x |F_n(x) - F(x)| \xrightarrow{\text{as}} 0.$$

- Kolmogorov-Smirnov theorem: under the same conditions,

$$\sqrt{n} \sup_x |F_n(x) - F(x)| \xrightarrow{\text{as}} \sup_{t \in [0,1]} |B(t)|,$$

where B is a Brownian bridge (standard Brownian motion subject to $B(0) = 0$ and $B(1) = 0$).

2 Probability: basic concentration inequalities

- Markov's inequality: if $X \geq 0$ and $\mu = \mathbb{E}[X]$, then for any $a > 0$,

$$\mathbb{P}(X \geq a) \leq \frac{\mu}{a}.$$

- Chebyshev's inequality: let $\mu = \mathbb{E}[X]$ and $\sigma^2 = \text{Var}[X]$, then for any $t > 0$,

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

- Hoeffding's inequality: if $X_1, \ldots, X_n$ are independent and mean zero with $a_i \leq X_i \leq b_i$ for each i, then for any $t > 0$,

$$\mathbb{P}(\bar{X}_n \geq t) \leq \exp\left(\frac{-2nt^2}{\frac{1}{n}\sum_{i=1}^n (b_i^2 - a_i^2)}\right).$$

- Bernstein's inequality: if $X_1, \ldots, X_n$ are independent and mean zero with $\sigma_i^2 = \text{Var}[X_i]$, and $|X_i| \leq M$ for each i, then for any $t > 0$,

$$\mathbb{P}(\bar{X}_n \geq t) \leq \exp\left(\frac{-nt^2/2}{\frac{1}{n}\sum_{i=1}^n \sigma_i^2 + Mt/3}\right).$$

3 Statistics: maximum likelihood and GLMs

3.1 Maximum likelihood

Let $\{P_\theta : \theta \in \Theta\}$ be a parametric family of models, and write p_θ for the probability density function (or probability mass function, in the discrete case). We write $L(\theta; z) = p_\theta(z)$ for the likelihood function.

Then given i.i.d. $Z_1, \ldots, Z_n \sim P_\theta$, the problem

$$\underset{\theta \in \Theta}{\text{maximize}} \prod_{i=1}^n L(\theta; Z_i),$$

is called *maximum likelihood estimation*, and a maximizer $\hat{\theta}$ of the above is called a *maximum likelihood estimator* (MLE). (We are often imprecise and call this *the* MLE, ignorning nonuniqueness of the solution.) Note that we can equivalently write this via the negative log-likelihood function $\ell(\theta; z) = -\log L(\theta; z)$, i.e., the above is equivalent to

$$\underset{\theta \in \Theta}{\text{minimize}} \sum_{i=1}^n \ell(\theta; Z_i).$$

The *Fisher information* is given by the expectation of the Hessian,

$$I(\theta) = \mathbb{E}[\nabla^2 \ell(\theta; Z)],$$

where the expectation is taken over $Z \sim P_\theta$. In the univariate case, $\Theta \subseteq \mathbb{R}$, this is simply

$$I(\theta) = \mathbb{E}\left[\frac{d^2\ell(\theta; Z)}{d\theta^2}\right].$$

Under appropriate regularity conditions, when $Z_1, \ldots, Z_n \sim P_\theta$,

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{s_n} \xrightarrow{d} N(0, 1),$$

where $s_n = 1/\sqrt{I(\hat{\theta})}$. In the multivariate case, $\Theta \subseteq \mathbb{R}^d$, under regularity conditions,

$$\sqrt{n} S_n^{-1}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I_d),$$

where $S_n = I(\hat{\theta})^{-1/2}$, which is the symmetric square root of the inverse Fisher information matrix at the MLE.

The following result is also true (under regularity conditions):

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I(\theta)^{-1}).$$

From this we can discern

$$\|\hat{\theta} - \theta\|_2^2 = O_p\left(\frac{\mathrm{tr}[I(\theta)^{-1}]}{n}\right),$$

and under standard conditions this will be $O_p(d/n)$. This should give you a rough sense for the squared error of the MLE, in standard (idealized) settings.[1] We will do a related but even simpler and more direct calculation later for least squares regression.

3.2 Generalized linear models

An *exponential family distribution* is of the form

$$p_\eta(z) = \exp\big(T(z)^\mathsf{T}\eta - \psi(\eta)\big)h(z),$$

Here $\eta \in \mathbb{R}^d$ is called the *natural parameter* of the exponential family (often we use η, rather than θ, here). We call the function ψ the *log-partition function*, we call T the *sufficient statistic*, and we call h the *base measure*.

A remarkable fact: in any exponential family distribution, the log-partition function ψ is always convex, by virtue of the fact that p_η must be a bona fide density and therefore must integrate to one (sum to one, in the discrete case). This means that the map $\eta \mapsto p_\eta(z)$ is always log-concave (for fixed z), and the resulting maximum likelihood problem

$$\underset{\eta}{\text{minimize}} \; -\frac{1}{n}\left(\sum_{i=1}^n T(Z_i)\right)^\mathsf{T}\eta + \psi(\eta)$$

is always a convex optimization problem. (Convexity and optimization will be discussed shortly.)

A *generalized linear model* (GLM) builds off the exponential family distribution. We observe independent draws of a response variable y_i, $i = 1, \ldots, n$, with each one sampled from an exponential family distribution. The form here is common (the functions T, ψ, h are common), but we have sample-specific natural parameters: to each y_i, we assign a natural parameter η_i, and model it as

$$\eta_i = x_i^\mathsf{T}\beta$$

[1] Note carefully, however, that for the asymptotic normal approximation to be accurate we require $d \ll n$; when d scales proportionally to n, $d/n \to \gamma > 0$, we get very different asymptotic behavior.

for a feature vector $x_i \in \mathbb{R}^d$ and parameter $\beta \in \mathbb{R}^d$. Maximum likelihood becomes:

$$\underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \Big(-T(y_i) x_i^\mathsf{T} \beta + \psi(x_i^\mathsf{T} \beta) \Big).$$

Maximum likelihood in a GLM has many special properties. For example, the Hessian of the criterion is the same as the Fisher information matrix,[2] since the second-order term has no dependence on y_i.

The three most important special cases are the Gaussian: $\psi(u) = \frac{u^2}{2}$, Bernoulli: $\psi(u) = \log(1 + e^u)$, and Poisson: $\psi(u) = e^u$ families. This gives rise to the following maximum likelihood problems,

$$\begin{aligned}
&\text{Gaussian :} && \underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \left(-y_i x_i^\mathsf{T} \beta + \frac{(x_i^\mathsf{T} \beta)^2}{2} \right) && \text{(where each } y_i \in \mathbb{R}) \\
&\text{Bernoulli :} && \underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \Big(-y_i x_i^\mathsf{T} \beta + \log(1 + \exp(x_i^\mathsf{T} \beta)) \Big) && \text{(where each } y_i \in \{0, 1\}) \\
&\text{Poisson :} && \underset{\beta}{\text{minimize}} \sum_{i=1}^{n} \Big(-y_i x_i^\mathsf{T} \beta + \exp(x_i^\mathsf{T} \beta) \Big) && \text{(where each } y_i \in \mathbb{N}).
\end{aligned}$$

These are known as *least squares regression*, *logistic regression*, and *Poisson regression*, respectively. The Hessians in these problems are:

$$\begin{aligned}
&\text{Gaussian :} && H(\beta) = \sum_{i=1}^{n} x_i x_i^\mathsf{T} && \text{(no dependence on } \beta) \\
&\text{Bernoulli :} && H(\beta) = \sum_{i=1}^{n} x_i p(x_i^\mathsf{T} \beta)(1 - p(x_i^\mathsf{T} \beta)) x_i^\mathsf{T} && \text{where } p(u) = 1/(1 + e^{-u}) \\
&\text{Poisson :} && H(\beta) = \sum_{i=1}^{n} x_i \lambda(x_i^\mathsf{T} \beta) x_i^\mathsf{T} && \text{where } \lambda(u) = e^u.
\end{aligned}$$

This can be written more succintly in terms of $X \in \mathbb{R}^{n \times d}$, the predictor matrix (whose i^{th} row is x_i):

$$\begin{aligned}
&\text{Gaussian :} && H(\beta) = X^\mathsf{T} X && \text{(no dependence on } \beta) \\
&\text{Bernoulli :} && H(\beta) = X^\mathsf{T} W(\beta) X && \text{where } W(\beta) \text{ is diagonal with } [W(\beta)]_{ii} = p(x_i^\mathsf{T} \beta)(1 - p(x_i^\mathsf{T} \beta)) \\
&\text{Poisson :} && H(\beta) = X^\mathsf{T} W(\beta) X && \text{where } W(\beta) \text{ is diagonal with } [W(\beta)]_{ii} = \lambda(x_i^\mathsf{T} \beta).
\end{aligned}$$

4 Convexity and optimization

A *convex set* $C \subseteq \mathbb{R}^d$ is one that satisfies

$$x, y \in C \implies tx + (1 - t)y \in C, \quad \text{for all } t \in [0, 1].$$

A *convex function* $f : \mathbb{R}^d \to (-\infty, \infty]$ is one such that its effective domain $\text{dom}(f)$ (the set of x for which f is defined and finite) is a convex set, and

$$f\big(tx + (1 - t)y\big) \leq tf(x) + (1 - t)f(y), \quad \text{for all } x, y \in \text{dom}(f) \text{ and } t \in [0, 1].$$

It is called *strictly convex* if the above inequality is satisfied strictly whenever $x \neq y$ and $t \in (0, 1)$. Concavity of f means that $-f$ is convex, and same with strict concavity.

A differentiable function f is convex if and only if $\text{dom}(f)$ is convex and

$$f(y) \geq f(x) + \nabla f(x)^\mathsf{T} (y - x), \quad \text{for all } x, y \in \text{dom}(f),$$

[2] Here we are referring to the total Fisher information matrix, which has been summed over the observations.

where ∇f denotes the gradient of f. This is called the *first-order characterization* of convexity. A twice differentiable function f is convex if and only if $\mathrm{dom}(f)$ is convex and

$$\nabla^2 f(x) \succeq 0, \quad \text{for all } x \in \mathrm{dom}(f).$$

where $\nabla^2 f$ denotes the Hessian of f, and we write $A \succeq 0$ to mean that a matrix A is positive semidefinite (symmetric and its smallest eigenvalue is nonnegative). This is called the *second-order characterization* of convexity.

From the expressions for the Hessians given above, you can check directly that the criterions in the least squares, logistic regression, and Poisson regression problems are convex.

4.1 Optimization basics

An *optimization problem* is of the form

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \; i = 1, \ldots, m \\ & h_j(x) = 0, \; j = 1, \ldots, k. \end{aligned}$$

Here the minimization is implicitly restricted to $D = \mathrm{dom}(f) \cap \bigcap_{i=1}^m \mathrm{dom}(g_i) \cap \bigcap_{j=1}^k \mathrm{dom}(h_j)$, the intersection of relevant effective domains.

A *convex optimization problem* is one of the above form such that f and g_i, $i = 1, \ldots, m$ are all convex functions, and h_j, $j = 1, \ldots, k$ are all affine functions. In other words, the problem

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \; i = 1, \ldots, m \\ & Ax = b, \end{aligned}$$

is convex whenever f and g_i, $i = 1, \ldots, m$ are convex (and A and b are arbitrary).

The function f in such problems is called the *objective* or *criterion*. A *feasible point* is a point in D such that all constraints (inequality and equality constraints) are met. The infimal criterion value among all feasible points is often denoted $f^\star$, and called the *optimal value*.

A feasible point that achieves the optimal value is denoted $x^\star$ (note that $f^\star = f(x^\star)$), and is called a *solution* or *minimizer*. However, in statistics, we often use "hat notation", as in $\hat{\theta}$ for the MLE. An important fact: if the criterion f is strictly convex, and a solution exists, then it must be unique.

A point $\bar{x}$ is called a *local solution* if it is feasible and there is some $\delta > 0$ such that

$$f(\bar{x}) \leq f(x), \quad \text{for all feasible } x \text{ such that } \|x - \bar{x}\|_2 \leq \delta.$$

For a convex optimization problem, the following holds: any local solution $\bar{x}$ must also be a global solution: $f(\bar{x}) \leq f(x)$ for all feasible points x. This result is so important that it may as well be called the *fundamental theorem of convex optimization*.

A huge number of estimators in statistics and machine learning are defined by optimization problems, and many of these are convex problems. There has been a surge in interest in (smooth) nonconvex optimization recently due to the rise of deep neural networks.

4.2 Subgradients

For a function f on $\mathbb{R}^d$, we say that $s \in \mathbb{R}^d$ is a *subgradient* of f at $x \in \mathrm{dom}(f)$ provided that

$$f(y) \geq f(x) + s^\mathsf{T}(y - x), \quad \text{for all } y \in \mathrm{dom}(f).$$

This is analogous to the first-order characterization for convexity, where s playes the role of $\nabla f(x)$. We write $\partial f(x)$ for the set of all subgradients of f at x, which is called the *subdifferential*.

For convex f, if f is differentiable at x, then $\partial f(x) = \{x\}$. The converse is true as well (if the subdifferential is a singleton, then f must be differentiable at x, with its gradient given by the single subgradient).

Subgradients play a big role in nonsmooth optimization, but are also key for statistical analysis for certain problems/estimators, such as the lasso. For example, the *subgradient optimality condition* gives us the following characterization:

$$x \text{ minimizes } f \iff 0 \in \partial f(x).$$

We note that this is true without any assumptions on f (no need to assume convexity of f). The proof is so simple it's nearly vacuous: x minimizes f if and only if $f(y) \geq f(x)$ for all y, which is the definition of 0 being a subgradient of f at x. And for differentiable convex f, we get the familiar condition $0 = \nabla f(x)$.

4.3 Algorithms

The two most basic methods for unconstrained minimization of a function f are *gradient descent*:

$$x_{k+1} = x_k - t_k \nabla f(x_k), \quad k = 1, 2, 3, \ldots,$$

for differentiable f, where each $t_k \geq 0$ is a step size, and *Newton's method*:

$$x_{k+1} = x_k - t_k [\nabla^2 f(x_k)]^{-1} \nabla f(x_k), \quad k = 1, 2, 3, \ldots,$$

for twice differentiable f, where again each $t_k \geq 0$ is a step size. For nonsmooth and/or constrained optimization, variants exist (as do many, many other optimization algorithms).

These two algorithms (or close cousins of them) appear frequently in statistics and machine learning. For example, for optimization in a GLM, Newton's method reduces to what is known as *iteratively reweighted least squares* (IRLS), which is important both computationally and inferentially.

5 Regression

Suppose that we observe (X, Y) from some unknown joint distribution, where $Y \in \mathbb{R}$, and we are interested in predicting Y from X. Over all functions f, the prediction error as measured in terms of squared loss

$$\mathbb{E}\big[(Y - f(X))^2\big]$$

is minimized at

$$f(x) = \mathbb{E}[Y|X = x],$$

called the *regression function* of Y on X. If we observe i.i.d. pairs (x_i, y_i), $i = 1, \ldots, n$ from the same joint distribution as (X, Y), then we can use this to estimate f. This is the most generic problem formulation available for regression, and we'll call it the *XY-Pairs model*.

5.1 Signal-plus-noise models

You'll often see authors writing down a model of the form:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n.$$

This is often referred to as a "signal-plus-noise" model. You might initially think that this is assuming more than an XY-Pairs model, but it's actually the same, for i.i.d. mean zero stochastic errors ϵ_i, $i = 1, \ldots, n$, provided we treat the predictors x_i, $i = 1, \ldots, n$ as suitably random as well. Let's ammend the above to emphasize this:

$$\begin{gathered} (x_i, y_i),\ i = 1, \ldots, n \text{ are i.i.d.,} \\ \text{where each } y_i = f(x_i) + \epsilon_i, \text{ and } \mathbb{E}[\epsilon_i] = 0. \end{gathered} \tag{1}$$

which we'll call the *Random-X signal-plus-noise model*. To see that

$$\text{XY-Pairs} \iff \text{Random-X signal-plus-noise},$$

the key is that we can always define $\epsilon_i = y_i - f(x_i)$. This has zero mean as $\mathbb{E}[f(x_i)] = \mathbb{E}[\mathbb{E}[y_i|x_i]] = \mathbb{E}[y_i]$.

But what if we assume that each x_i is fixed? To contrast, we'll call this the *Fixed-X signal-plus-noise model*:

$$\begin{gathered} x_i,\ i = 1, \ldots, n \text{ are fixed}, \\ \delta_i,\ i = 1, \ldots, n \text{ are i.i.d.}, \\ \text{where each } y_i = f_0(x_i) + \delta_i, \text{ and } \mathbb{E}[\delta_i] = 0. \end{gathered} \tag{2}$$

The interpretation you'll often hear: this is just given by the Random-X signal-plus-noise model (1) (which is itself equivalent to the XY-Pairs model) after we condition on each x_i.

However, there is an important subtlety here! The precise connection is this:

$$\begin{aligned} \text{Fixed-X signal-plus-noise} \iff \text{Conditioning on each } x_i \text{ in Random-X signal-plus-noise,} \\ \textit{provided that each } x_i \perp\!\!\!\perp \epsilon. \end{aligned}$$

The reason for the extra independence assumption is this: if didn't have independence of each x_i and ϵ_i, then after conditioning on x_i, $i = 1, \ldots, n$ in (1), the distributions of $\delta_i = \epsilon_i | x_i$, $i = 1, \ldots, n$ need not be identical.

5.2 Independence of x_i and ϵ_i

To be frank, assuming independence of x_i and ϵ_i is fairly common. But to be clear, it is an *extra assumption* and does not fall out of the generic XY-Pairs model. It's also not totally harmless. To see this, suppose that we were in a situation where the true model is:

$$\begin{gathered} (x_i, z_i, y_i),\ i = 1, \ldots, n \text{ are i.i.d.}, \\ \text{where each } y_i = f(x_i) + g(z_i) + \xi_i, \text{ and } \mathbb{E}[\xi_i] = 0, \end{gathered}$$

Let's even assume that each $x_i \perp\!\!\!\perp \xi_i$, and $\mathbb{E}[g(z_i)|x_i] = 0$. Then the regression function (of y_i on x_i) is indeed f. But in the signal-plus-noise model for y_i on x_i, the noise variable is

$$\epsilon_i = g(z_i) + \xi_i,$$

and if z_i is dependent on x_i, then ϵ_i will be too.

Therefore, assuming that ϵ_i is independent of x_i, in general, is like assuming that *any omitted variables are independent of the current ones*. Which could certainly be seen as a strong assumption.

5.3 Prediction error metrics

The Random-X and Fixed-X settings call to mind similar but distinct notions of prediction error. In the former, it is natural to ask: how would well would we predict at a new test point (x_0, y_0)? That is, given an estimator $\hat{f}$ that we fit to training data (x_i, y_i), $i = 1, \ldots, n$, we consider

$$\text{Err}(\hat{f}) = \mathbb{E}\big[(y_0 - \hat{f}(x_0))^2\big], \tag{3}$$

where the expectation is taken over the training data (x_i, y_i), $i = 1, \ldots, n$ and the test point (x_0, y_0), all i.i.d. We'll call $\text{Err}(\hat{f})$ the *out-of-sample prediction error*, often (here and frequently with other terms as well) dropping "prediction" for simplicity.

Meanwhile, in the Fixed-X setting, it would instead be more natural to consider

$$\text{Err}(\hat{f}; x_{1:n}) = \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^{n}(y_i^* - \hat{f}(x_i))^2\bigg], \tag{4}$$

where the expectation is taken each over y_i and an independent copy y_i^* of the response value. The notation $\mathrm{Err}(\hat{f}; x_{1:n})$ emphasizes the dependence of this metric, which we call the *in-sample prediction error*,[3] on $x_{1:n} = \{x_1, \ldots, x_n\}$. It can be interpreted as the prediction error *had we encountered new responses at the exact same feature values.*

If you think the latter interpretation sounds strange, then—well, you're right, it is a bit strange in a modern prediction context! (More discussion soon.)

5.3.1 Risk equivalents

Assuming independence of $\epsilon_0 = y_0 - f(x_0)$ and x_0, we have, by adding and subtracting $f(x_0)$ in (3) and expanding:

$$\underbrace{\mathbb{E}\big[(y_0 - \hat{f}(x_0))^2\big]}_{\mathrm{Err}(\hat{f})} = \underbrace{\mathbb{E}\big[(f(x_0) - \hat{f}(x_0))^2\big]}_{\mathrm{Risk}(\hat{f})} + \sigma^2,$$

where $\sigma^2 = \mathrm{Var}[\epsilon_0]$ is the noise variance (also known as the *Bayes error* or *irreducible error* in the current context). We'll call the first term on the right-hand side the *out-of-sample prediction risk*[4] and denote it by $\mathrm{Risk}(\hat{f})$. Like the above relationship, we can also write down a similar equivalence to risk in the Fixed-X context, by adding and subtracting $f(x_i)$ in each summand of (4) and expanding:

$$\underbrace{\mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n (y_i^* - \hat{f}(x_i))^2\bigg]}_{\mathrm{Err}(\hat{f}; x_{1:n})} = \underbrace{\mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n (f(x_i) - \hat{f}(x_i))^2\bigg]}_{\mathrm{Risk}(\hat{f}; x_{1:n})} + \sigma^2,$$

where $\sigma^2 = \mathrm{Var}[\epsilon_i]$ is the noise variance. Similarly, we'll call the first term on the right-hand side the *in-sample prediction risk*, and denote it by $\mathrm{Risk}(\hat{f}; x_{1:n})$.

So to recap, in the Random-X setting, we have as a natural metric

$$\mathrm{Risk}(\hat{f}) = \mathbb{E}\big[(f(x_0) - \hat{f}(x_0))^2\big], \tag{5}$$

called the out-of-sample risk, which only differs from the out-of-sample error by σ^2. Meanwhile, in the Fixed-X setting, we have the natural metric

$$\mathrm{Risk}(\hat{f}; x_{1:n}) = \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n (f(x_i) - \hat{f}(x_i))^2\bigg], \tag{6}$$

called the in-sample risk, which again only difference from the in-sample error by σ^2.

5.3.2 Interlude: same roses, different names

Note that we can also view $\mathrm{Risk}(\hat{f})$ and $\mathrm{Risk}(\hat{f}; x_{1:n})$ in terms of the $L^2(P)$ and $L^2(P_n)$ error metrics:

$$\mathrm{Risk}(\hat{f}) = \mathbb{E}\big\|f - \hat{f}\big\|_{L^2(P)}^2 = \mathbb{E}\bigg[\int (f(x) - \hat{f}(x))^2 \, dP(x)\bigg],$$

$$\mathrm{Risk}(\hat{f}; x_{1:n}) = \mathbb{E}\big\|f - \hat{f}\big\|_{L^2(P_n)}^2 = \mathbb{E}\bigg[\int (f(x) - \hat{f}(x))^2 \, dP_n(x)\bigg],$$

where P is the feature distribution and P_n the empirical distribution of $x_{1:n}$. Thus out-of-sample and in-sample error could also be called the (expected) $L^2(P)$ *risk* and $L^2(P_n)$ *risk*, respectively. Other names you'll hear are the *population* L^2 *risk* and *empirical* L^2 *risk*, respectively.

Empirical process theory gives us bounds on the $L^2(P_n)$ and $L^2(P)$ norms (typically, these bounds hold for all functions in some particular function class, with high probability over draws of $x_1, \ldots, x_n$). We'll cover this later in the course.

[3] Don't confuse this with training error! This may be tempting as the name "in-sample" may remind you of training error. But they're not the same thing! (Even worse, some other authors use "in-sample error" and "training error" synonomously).

[4] Admittedly, many authors use the terms "risk" and "error" interchangeably, and do not use it to distinguish whether the target is $f(x_0)$ or y_0. In fact, we will do so ourselves in future lectures! However, when considering both metrics in a single discussion—either (3) and (5), or (4) and (6)—it is helpful to use nomenclature to distinguish the two.

5.3.3 Discussion: how different are they?

Classically, statisticians really love in-sample risk (6). A big part of this is probably that it gives a very clean framework in which we can develop various powerful tools for analysis (e.g., the covariance formula, Stein's unbiased risk estimator, etc.). To be more charitable, there may have also been more genuine motivations for in-sample risk from experimental design: if the values $x_1, \dots, x_n$ were chosen (i.e., designed), and we ran an experiment in which we measured $y_1, \dots, y_n$, then looking at $\mathrm{Err}(\hat{f}; x_{1:n})$ would not seem so crazy.

In machine learning and in most modern application of statistical prediction, however, out-of-sample risk (5) seems to be much more fitting. So this begs the question: how different can (5) and (6) be? This turns out to be a rich question, and there is a lot to say about it. The short answer is:

- *usually not very different* in classical regimes (low-dimensional, smooth $\hat{f}, f$); but
- *can be very different* in modern regimes (high-dimensional, and/or nonsmooth $\hat{f}, f$).

For example, in the interpolation regime (of great interest recently, along with overparametrized machine learning methods more generally), they can be extremely different. Again, we may study this later in the course if there is time and interest. For now, we'll just look at least squares regression which already provides some insights into the matter. Before that, we'll quickly review the bias-variance decomposition.

5.4 Bias and variance

In general, for an estimator $\hat{\theta}$ of θ, it holds (just add and subtract $\mathbb{E}[\hat{\theta}]$, and expand) that

$$\mathbb{E}[(\theta - \hat{\theta})^2] = \underbrace{(\theta - \mathbb{E}[\hat{\theta}])^2}_{\mathrm{Bias}^2(\hat{\theta})} + \underbrace{\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]}_{\mathrm{Var}(\hat{\theta})}.$$

This is commonly known as the *bias-variance decomposition*. In a Fixed-X regression context, where the feature values $x_{1:n} = \{x_1, \dots, x_n\}$ are treated as fixed, we just apply this to each $\hat{\theta} = \hat{f}(x_i)$ and average over $i = 1, \dots, n$ to obtain

$$\mathrm{Risk}(\hat{f}; x_{1:n}) = \frac{1}{n} \sum_{i=1}^n \mathrm{Bias}^2(\hat{f}(x_i)) + \frac{1}{n} \sum_{i=1}^n \mathrm{Var}(\hat{f}(x_i)).$$

Meanwhile, in a Random-X setting, we can condition on the test point x_0, apply this to $\hat{\theta} = \hat{f}(x_0)$, and then integrate over x_0 to obtain

$$\mathrm{Risk}(\hat{f}) = \mathbb{E}\big[\mathrm{Bias}^2(\hat{f}(x_0)|x_0)\big] + \mathbb{E}\big[\,\mathrm{Var}(\hat{f}(x_0)|x_0)\big],$$

where the expectation is over x_0.

The bias-variance decomposition is useful for a variety of reasons, from analytical to conceptual. The typical trend is that underfitting means high bias and low variance, whereas overfitting means low bias but high variance. And the conventional wisdom is that we want to balance these in order to make accurate predictions.

Interestingly, the bias-variance decomposition has been called into question in recent years, with regards to the study of overparametrized estimators in machine learning. But I don't really think that the fundamental idea of "balancing bias and variance being a good thing" is actually contradicted here. It's more that bias and variance can manifest themselves in strange ways in these settings. We'll cover this later in the course.

Finally, it should be noted that the decompositions presented above aren't the only bias-variance decompositions available, and certainly not the only ones that are useful. We can obtain other ones by conditioning on—and later integrating out over—other parts of data. For example, in the next subsection, to analyze the risk of least squares, we will see that it is useful to also condition on the training features.

5.5 Least squares regression

Let $X \in \mathbb{R}^{n \times d}$ denote the predictor matrix (whose i^{th} row is x_i). Analogously, let $Y \in \mathbb{R}^n$ denote the vector of response variables. *Least squares regression* of Y on X is given by

$$\hat{f}(x) = x^\mathsf{T} \hat{\beta},$$

where

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \ \|Y - X\beta\|_2^2 = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y.$$

Critically, here we are assuming that $\operatorname{rank}(X) = d$, which necessarily requires that $d \leq n$, which makes the least squares criterion admit a unique solution. Note that the fitted values (i.e., in-sample predictions) are

$$X\hat{\beta} = X(X^\mathsf{T} X)^{-1} X^\mathsf{T} Y = P_X Y,$$

where $P_X = X(X^\mathsf{T} X)^{-1} X^\mathsf{T}$ denotes the projection onto the column space of X.

5.5.1 In-sample risk

To investigate its risk properties with as simple math as possible, let's assume an underlying linear model

$$Y = X\beta_0 + \epsilon, \tag{7}$$

with X fixed, and $\epsilon \in \mathbb{R}^n$ having i.i.d. entires with mean zero and variance σ^2. First, observe that

$$\mathbb{E}[X\hat{\beta}] = P_X \mathbb{E}[Y] = P_X X \beta_0 = X\beta_0,$$

so least squares regression has zero bias, in the in-sample sense. Thus its in-sample error is pure variance, and we compute

$$\operatorname{Risk}(\hat{f}; x_{1:n}) = \frac{1}{n} \operatorname{tr}[\operatorname{Var}(P_X Y)] = \sigma^2 \frac{d}{n}. \tag{8}$$

5.5.2 Out-of-sample risk

Meanwhile, for the out-of-sample risk we'll take X to be random, and assume X is independent of ϵ in (7), with (x_0, y_0) being another i.i.d. draw from the same linear model. Then, conditioning on both X, x_0, we can see that

$$\mathbb{E}[x_0^\mathsf{T} \hat{\beta} \,|\, X, x_0] = x_0^\mathsf{T} (X^\mathsf{T} X)^{-1} X^\mathsf{T} \mathbb{E}[Y|X] = x_0^\mathsf{T} \beta_0.$$

Hence the out-of-sample bias will still be zero, after integrating over X, x_0. (Note that unbiasedness here doesn't actually require the strong assumption of X, ϵ being independent.) To compute the out-of-sample variance, we again first condition on X, x_0:

$$\operatorname{Var}(x_0^\mathsf{T} \hat{\beta} \,|\, X, x_0) = x_0^\mathsf{T} (X^\mathsf{T} X)^{-1} X^\mathsf{T} \operatorname{Var}(Y|X) X (X^\mathsf{T} X)^{-1} x_0 = \sigma^2 x_0^\mathsf{T} (X^\mathsf{T} X)^{-1} x_0.$$

Then integrating over X, x_0 gives the out-of-sample risk:

$$\operatorname{Risk}(\hat{f}) = \sigma^2 \operatorname{tr}\Big(\mathbb{E}[x_0 x_0^\mathsf{T}] \, \mathbb{E}[(X^\mathsf{T} X)^{-1}]\Big), \tag{9}$$

where we have used the independence of X, x_0. An exact formula will not be possible in full generality here, since as we can see the out-of-sample risk depends on the distribution of the predictors. Contrast this with the in-sample risk, which did not.

However, we can still go further than (9) from several perspectives.

- In the special case when each $x_i \sim N(0, \Sigma)$, with Σ invertible, we can compute the out-of-sample risk exactly. In this case, it holds that $X^\mathsf{T} X \sim W(\Sigma, n)$, a Wishart distribution, and $(X^\mathsf{T} X)^{-1} \sim W^{-1}(\Sigma^{-1}, n)$, an inverse Wishart distribution. Thus

$$\operatorname{Risk}(\hat{f}) = \sigma^2 \operatorname{tr}\left(\Sigma \frac{\Sigma^{-1}}{n - d - 1}\right) = \sigma^2 \frac{d}{n - d - 1}.$$

- If we assume conditions that are standard in random matrix theory (which permit Gaussian features but also many other types of feature distributions), and let $d/n = \gamma \in (0, 1)$, then as d, n grow,

$$\text{Risk}(\hat{f}) \approx \sigma^2 \frac{\gamma}{1-\gamma}.$$

  This agrees with the normal calculation from the last bullet point. Comparing the above display to (8), which we can simply write as $\text{Risk}(\hat{f}; x_{1:n}) = \sigma^2\gamma$, we see that the in-sample and out-of-sample risks can be very different—with the latter being much larger—when γ is close to 1 (d is close to n). We'll revisit this and related calculations later in the course.

- In general, we can always rewrite (9) as

$$\text{Risk}(\hat{f}) = \frac{\sigma^2}{n} \text{tr}\left(\mathbb{E}[X^\mathsf{T}X]\,\mathbb{E}[(X^\mathsf{T}X)^{-1}]\right),$$

  just by using the fact that x_0 and the rows of X have the same distribution. It can be shown (Groves and Rothenberg, 1969) that

$$\mathbb{E}[(X^\mathsf{T}X)^{-1}] - (\mathbb{E}[X^\mathsf{T}X])^{-1} \text{ is positive semidefinite,}$$

  for any feature distribution such that $X^\mathsf{T}X$ is almost surely invertible. Then from the above display,

$$\text{Risk}(\hat{f}) \geq \frac{\sigma^2}{n} \text{tr}\left(\mathbb{E}[X^\mathsf{T}X]\,(\mathbb{E}[X^\mathsf{T}X])^{-1}\right) = \sigma^2 \frac{d}{n}.$$

  That is, the out-of-sample risk is always larger than the in-sample risk for least squares.

  On the homework, you will generalize this last result by showing that it still holds without assuming a true linear model (i.e., allowing the regression function to be possibly nonlinear).

5.6 Regularization

As we just saw, the risk of least squares regression degrades as d grows close to n—in the case of Gaussian features or "RMT features" (which we use to mean feature models compatible with the standard assumptions in random matrix theory), the out-of-sample risk actually diverges when $d = n$. Meanwhile, the least squares estimator is not even well-defined when $d > n$, in that the optimization problem

$$\underset{\beta}{\text{minimize}} \ \|Y - X\beta\|_2^2$$

does not have a unique solution.

How do we deal with such issues? The short answer is *regularization*. At its core, regularization provides us with a way to navigate the bias-variance tradeoff: we (ideally greatly) reduce the variance at the expense of introducing bias. In regression, canonical choices of regularizers are the ℓ_0, ℓ_1, and ℓ_2 norms:

$$\|\beta\|_0 = \sum_{j=1}^d 1\{\beta_j \neq 0\}, \quad \|\beta\|_1 = \sum_{j=1}^d |\beta_j|, \quad \|\beta\|_2 = \left(\sum_{j=1}^d \beta_j^2\right)^{1/2}.$$

This gives rise to the regularized least squares problems:

$$\text{Best subset selection :} \quad \underset{\beta}{\text{minimize}} \ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_0 \tag{10}$$

$$\text{Lasso regression :} \quad \underset{\beta}{\text{minimize}} \ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \tag{11}$$

$$\text{Ridge regression :} \quad \underset{\beta}{\text{minimize}} \ \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2, \tag{12}$$

where $\lambda \geq 0$ is called the tuning parameter (typically chosen by cross-validation or similar techniques.)

One of these is not like the other: convexity. Calling $\|\cdot\|_0$ the "ℓ_0 norm" is a misnomer, since it is not a norm. Critically, $\|\cdot\|_0$ is *not convex*, while $\|\cdot\|_1$ and $\|\cdot\|_2$ are (indeed, any norm is a convex function). This makes (10) a nonconvex problem, and one that is generally very hard to solve in practice except for very small d (e.g., its constrained form is known to be NP-hard). On the other hand, problems (11) and (12) are convex optimization problems, and many efficient algorithms exist for them.

One of these is not like the other: sparsity. Meanwhile, best subset selection and the lasso have a special and useful property: their solutions $\hat{\beta}$ are *sparse*, which means $\hat{\beta}_j = 0$, for many j. Larger λ typically means sparser solutions. This is not true of the ridge regression estimator, which will be generically dense (all nonzero components), for any $\lambda \geq 0$. Sparsity is often desirable because it corresponds to performing variable selection in the fitted linear model.

6 Classification

We conclude with a quick recap of classification. Many of the same ideas from regression carry over. For $(X, Y) \sim P$, with $Y \in \{0, 1\}$, the regression function is

$$f(x) = \mathbb{E}[Y|X = x] = \mathbb{P}(Y = 1|X = x),$$

which now becomes the conditional probability of observing class 1, given $X = x$. Over all classifiers C, the one that minimizes misclassification risk

$$\text{Risk}(C) = \mathbb{P}(Y \neq C(X))$$

is called the *Bayes classifier*, defined by

$$C(x) = \begin{cases} 0 & \text{if } f(x) \leq 1/2 \\ 1 & \text{if } f(x) > 1/2 \end{cases}.$$

To see this, consider any classifier C' and any fixed x,

$$\begin{aligned} \mathbb{P}(Y \neq C'(X) \,|\, X = x) &= 1 - \mathbb{P}(Y = 1, C'(X) = 1 \,|\, X = x) - \mathbb{P}(Y = 0, C'(X) = 0 \,|\, X = x) \\ &= 1 - C'(x)f(x) + (1 - C'(x))(1 - f(x)) \\ &= f(x) + (1 - 2f(x))C'(x). \end{aligned}$$

Thus, we can compute the conditional risk difference between C' and C as

$$\mathbb{P}(Y \neq C'(X) \,|\, X = x) - \mathbb{P}(Y \neq C(X) \,|\, X = x) = (2f(x) - 1)(C(x) - C'(x)).$$

When $f(x) > 1/2$, we have $C(x) = 1$ by construction, and so the right-hand side above is nonnegative. When $f(x) \leq 1/2$, we have $C(x) = 0$ by construction, and so again the right-hand side is nonnegative. Therefore we have shown $\mathbb{P}(Y \neq C'(X)|X = x) - \mathbb{P}((Y \neq C(X)|X = x) \geq 0$ for every x; integrating over x gives the result

$$\mathbb{P}(Y \neq C(X)) \leq \mathbb{P}(Y \neq C'(X)),$$

for any classifier C'.

Many classifiers are *plug-in classifiers*, of the form

$$\hat{C}(x) = \begin{cases} 0 & \text{if } \hat{f}(x) \leq 1/2 \\ 1 & \text{if } \hat{f}(x) > 1/2 \end{cases},$$

for an estimator $\hat{f}$ of the regression function. It is often useful to recode so that $Y \in \{-1, 1\}$, because then plug-in classifiers take the form

$$\hat{C}(x) = \text{sign}(\hat{h}(x)),$$

for some $\hat{h}$.

Note that in this coding, we can write misclassification risk as

$$\text{Risk}(\hat{C}) = \mathbb{P}(Y \neq \hat{C}(X)) = \mathbb{P}(Y\hat{h}(X) < 0).$$

This is of the form

$$\mathbb{E}[L(Y\hat{h}(X))], \quad \text{where } L(u) = 1\{u < 0\},$$

where L is clearly a nonconvex function. Often, we replace this with a convex surrogate loss, such as

$$\begin{aligned} &\text{Logistic :} && L(u) = \log(1 + \exp(-u)) \\ &\text{Exponential :} && L(u) = \exp(-u) \\ &\text{Hinge :} && L(u) = (1 - u)_+, \end{aligned}$$

as used in logistic regression, AdaBoost, and support vector machines, respectively.

References

Theodore Groves and Thomas Rothenberg. A note on the expected value of an inverse matrix. *Biometrika*, 56(3):690–691, 1969.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2009. Second edition.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004.