01 Intro & Regression Basics

1 Prereqs/background

Much of 36-708 will involve learning not only how an algorithm/method works operationally, but also how/whether it works in terms of error/accuracy, and when, i.e., under what assumptions. To study these kinds of questions, we will need to rely on a shared language/background of probability and statistics.

In particular, 36-705 (Intermediate Statistics) is an important prerequisite for this course. The following gives a brief list of some fundamental concepts; if anything is unclear, you should make sure to brush up and/or consider other courses before taking 708.

- concentration inequalities
- modes of convergence
- the central limit theorem
- empirical processes / uniform laws
- method of moments / MLE
- Cramer-Rao lower bounds
- minimax rates
- hypothesis testing
- confidence sets
- bootstrap
- Gaussian sequence model
- model selection

It is assumed you have seen linear regression in both the low- and high-dimensional setting, as well as some nonparametric regression methods (k-nearest neighbors, kernel methods). These will be covered at the start of the course as building blocks for what follows.

2 What we will cover this semester

The course 36-705 gave you background in mathematical statistics at an intermediate level, and in 36-707 you got some experience with basic regression methods and applying them in data analyses. This class will build off both of these previous courses, to study some modern statistical machine learning methods, and gain an undertstanding of the theory underlying the methods. The high-level mission of this course is to give you a good roadmap of different goals in ML, what methods can be used to achieve these goals, and what guarantees the methods can provide, under what conditions. It will therefore be a mix of theory, methods, and application. As detailed in the syllabus, some planned topics for the course include:

- nonparametric regression
- density estimation
- generalized additive models
- random forests
- neural networks
- high-dimensional regression
- cross-validation
- aggregation
- semisupervised learning
- classification
- conformal prediction
- inference
- clustering
- causality
- functional estimation
- minimax optimality

3 Statistics and machine learning

Statistics and ML have lots of overlap, and center on how to extract information from data. From Mohri et al: "Machine learning can be broadly defined as computational methods using [past information] to improve performance or to make accurate predictions."

However, they have different emphases, both in academic focus and culture. For example, the following topics often have more emphasis in ML:

- bandits, reinforcement learning, computational complexity, deep learning

While here are some topics that tend to have more emphasis in statistics:

- confidence sets, large sample theory, minimax optimality, causality

Topics like prediction, probability bounds (concentration), clustering, etc. tend to be commonly studied in both areas. This is just a small sampling, though, and these things vary by group, over time, etc. For example, RKHS methods started in statistics and are now probably more popular in ML circles.

Some cultural differences are that ML researchers tend to publish shorter papers in conferences, while statisticians tend to publish longer papers in journals. There is also a difference in notation and terminology. Here is an example from Wasserman's "All of statistics" textbook:

Statistics/Data Mining Dictionary

Statisticians and computer scientists often use different language for the same thing. Here is a dictionary that the reader may want to return to throughout the course.

Statistics	Computer Science	Meaning
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete Y from X
clustering	unsupervised learning	putting data into groups
data	training sample	$(X_1, Y_1), \ldots, (X_n, Y_n)$
covariates	features	the X_i's
classifier	hypothesis	a map from covariates to outcomes
hypothesis	—	subset of a parameter space Θ
confidence interval	—	interval that contains an unknown quantity with given frequency
directed acyclic graph	Bayes net	multivariate distribution with given conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update beliefs
frequentist inference	—	statistical methods with guaranteed frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

However, it seems the fields are overlapping more and more each year.

4 Regression basics

Regression is the problem of estimating the conditional expectation function

$$\mu(x) = \mathbb{E}(Y \mid X = x)$$

from samples $(X_1, Y_1), ..., (X_n, Y_n)$. Here the variables $Y \in \mathbb{R}$ are called outcomes (or sometimes labels if discrete) and the variables $X \in \mathbb{R}^d$ are called covariates or features. We call μ either the conditional expectation or regression function.

4.1 Why regression?

Regression is one of the most fundamental tasks in all of machine learning, and there are many reasons why it matters. First, one may be specifically interested in how outcomes Y are related to features X, and a conditional expectation is a reasonable starting place to summarize that relationship; in this case our goal is really to estimate the regression function μ. Alternatively, we may be interested in estimating a regression function solely for a subsequent downstream task, such as making predictions, doing clustering, summarizing variable importance, making causal inferences, etc.

One way to see why regression matters for prediction, in particular, is as follows. Suppose we want to construct a function $h : \mathcal{X} \mapsto \mathcal{Y}$ for predicting/guessing outcomes Y from features X. A natural way to measure the quality of our function for this task is through the risk

$$R(h) = \mathbb{E}\Big[\{Y - h(X)\}^2\Big]$$

which characterizes how far our guess $h(X)$ is from Y, in terms of squared error, in expectation. Then it follows that

$$\arg\min_h \mathbb{E}\Big[\{Y - h(X)\}^2\Big] = \mu$$

i.e., the function that that does best in terms of predicting outcomes, with respect to this least squares risk, is exactly the regression function, which looks at the features $X = x$ and spits out the conditional expectation $\mu(x) = \mathbb{E}(Y \mid X = x)$. This follows from the fact that

$$\begin{aligned}
\mathbb{E}\Big[\{Y - h(X)\}^2\Big] &= \mathbb{E}\Big[\{Y - \mu(X) + \mu(X) - h(X)\}^2\Big] \\
&= \mathbb{E}\Big[\{Y - \mu(X)\}^2\Big] + \mathbb{E}\Big[\{\mu(X) - h(X)\}^2\Big] \\
&\geq \mathbb{E}\Big[\{Y - \mu(X)\}^2\Big]
\end{aligned}$$

where the second line follows by iterated expectation, and the minimum is indeed attained when $h(X) = \mu(X)$ almost everywhere. We will see many more examples of downstream tasks where regression plays a crucial role later.

4.2 Sampling model

For the sampling model of the observations $(X_1, Y_1), ..., (X_n, Y_n)$ we will either assume:

1. the observations (X_i, Y_i) are iid, drawn from some joint distribution $\mathbb{P}$ of (X, Y), or

2. the outcomes Y_i are conditionally independent given $X^n \equiv (X_1, ..., X_n)$, and Y_i is conditionally independent of $X^n_{-i} = X^n \setminus X_i$ given X_i, i.e., the likelihood is

$$p(y_1, ..., y_n \mid x_1, ..., x_n)p(x_1, ..., x_n) = \prod_{i=1}^{n} p(y_i \mid x_i)p(x_1, ..., x_n).$$

Note that (1) implies (2), and assumption (2) is weaker since no independence assumptions are made about the features X^n: they could be arbitrarily dependent, or even non-random. Setting (1) is typically called "random design" and setting (2) is called "fixed design". In (2) results are often framed conditionally given all the covariates X^n.

Usually, estimators do not differ much if at all in the two designs (i.e., one uses the same methods in either case), so that one can derive results under the fixed design (2) and then obtain results for the random design (1) for free. However, we will see that results can differ slightly in the two settings, even analyzing the same estimator, in some cases.

Which model one uses could be motivated by generalization. Do we believe each feature observation X_i is a representative draw from some marginal distribution, and do we want to generalize from our sample to a larger population? Or do we think we have a haphazard unique sample, and only restrict our inferences to what happened for these particular units?

Remark 1. Often one sees authors additionally imposing a constant variance assumption, which says that $\text{var}(Y \mid X = x) = \sigma^2$ does not vary with features X. We will avoid this.

4.3 Measuring accuracy

Suppose we have a particular estimator $\widehat{\mu}(x)$ of the regression function $\mu(x)$. How do we say whether or how good the estimator is? There are potentially many measures of risk/accuracy, which should ideally reflect the underlying goal of the analysis. Often risk is measured via the expectation of some convex loss function, e.g., squared error. Some common examples are pointwise error

$$\mathbb{E}|\widehat{\mu}(x) - \mu(x)|,$$

L_2 error

$$\mathbb{E} \int \Big\{\widehat{\mu}(x) - \mu(x)\Big\}^2 \, d\mathbb{P}(x) = \mathbb{E}\|\widehat{\mu} - \mu\|^2,$$

uniform or ℓ_∞ error

$$\mathbb{E}\Big\{ \sup_x |\widehat{\mu}(x) - \mu(x)|\Big\} = \mathbb{E}\|\widehat{\mu} - \mu\|_\infty,$$

etc. Gyorfi et al. make the argument that L_2 risk is most natural in prediction-based regression problems, since

$$\mathbb{E}\Big[\{Y - \widehat{\mu}(X)\}^2\Big] = \mathbb{E}\Big[\{Y - \mu(X)\}^2\Big] + \mathbb{E} \int \Big\{\widehat{\mu}(x) - \mu(x)\Big\}^2 \, d\mathbb{P}(x)$$

and the first term is an unavoidable error just depending on inherent noise. However the choice of risk function is not always obvious and may depend closely on the specific goals of

the problem at hand.

The above notions of risk are all pointwise in that they depend on a particular distribution $\mathbb{P}$. It is also common to consider uniform worst-case versions of risk, e.g.,

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P \|\widehat{\mu} - \mu_P\|$$

which look at the worst possible risk over all distributions in some model $\mathcal{P}$ (recall a model is just a set of distributions, which we typically assume contains the true distribution).

4.4 No free lunch

In previous courses you have explored estimation and inference under a parametric model

$$\mu(x) = m(x; \beta)$$

for some finite-dimensional $\beta \in \mathbb{R}^p$. For example, in linear regression one assumes $\mu(x) = x^{\mathrm{T}}\beta$, in a generalized linear model one assumes $\mu(x) = g(x^{\mathrm{T}}\beta)$, e.g., $g(t) = \mathrm{expit}(t)$ in logistic regression, etc. You have also learned how to construct and analyze estimators in such cases using method of moments or maximum likelihood, for example.

Of course, in real life these kinds of parametric assumptions may be unlikely to hold exactly. One option then is to use the same methods, but resort to estimating a model-based summary/approximation of the regression function; however this could be quite different from the actual regression function, and in many cases we want to estimate the actual conditional expectation, not some summary.

This raises the question of how well one can do estimating the regression function without these parametric assumptions. Unfortunately, we can do arbitrarily poorly if we do not impose any non-trivial structure. The following results illustrate this fact (we will go through these results, including their proofs, in more detail later).

Theorem 1 (Györfi et al. [2002], Theorem 3.1)**.** *For any regression estimator $\widehat{\mu}(x)$, there exists a distribution of (X, Y) satisfying:*

1. *X is uniformly distributed,*
2. *$Y = \mu(X) \in \{-1, 1\}$ (i.e., Y is noiseless), and*

$$\limsup_{n \to \infty} \frac{\mathbb{E} \int \{\widehat{\mu}(x) - \mu(x)\}^2 \, dx}{1/\log n} \geq 1.$$

This shows that for any regression estimator, there exists a very simple but very difficult distribution, in the sense that integrated mean squared error rates can be no better than logarithmic in n (in fact Györfi et al. [2002] show that the above holds for any sequence $a_n \to 0$, logarithmic or not, and that a finite-sample version without the lim sup holds as well). Thus if we do not rule out such distributions somehow, regression can be an arbitrarily difficult statistical problem.

Theorem 2 (Tsybakov [2009], Theorem 2.8)**.** *Let $\mathcal{P}$ denote the set of distributions for which:*

1. *X has a bounded density on $[0,1]^d$, and*
2. *$\mu(x)$ is Lipschitz of order $s > 0$, i.e., $|\mu(x) - \mu(x')| \leq L\|x - x'\|^s$.*

Then

$$\inf_{\widehat{\mu}} \sup_{P \in \mathcal{P}} \mathbb{E}_P \int \left\{\widehat{\mu}(x) - \mu(x)\right\}^2 dP(x) \;\geq\; C n^{\frac{-1}{1+d/2s}}$$

for n large enough, where C is some constant not depending on n.

This shows that when the regression function is not smooth enough and/or when feature dimension is high (i.e., when s is very small or d is very large), no estimator can perform uniformly well, even if we only consider Lipschitz continuous regression functions. For example, suppose $d = 10$ and $s = 1/10$ then the best rate of convergence is $n^{-1/51}$, which is very slow. Even if $s = 1$ but dimension d is large, the rate can also be slow; this is the curse of dimensionality.

The above results make it clear that some structure will need to be imposed in order to have any hope at estimating a regression function with non-trivial accuracy. This raises important questions about what kinds of structure should be considered, and how to attain the best possible error rates under such structures.

References

L. Györfi, M. Kohler, A. Krzykaz, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer, 2002.

A. B. Tsybakov. *Introduction to Nonparametric Estimation.* New York: Springer, 2009.