10/36-702 Review and Introduction

These are things you should know from 36-705 and 10-715. Plus I will introduce a few new topics that we will cover in more detail later in the course.

1 Probability

1. $X_n \xrightarrow{P} 0$ means that means that, for every $\epsilon > 0$ $\mathbb{P}(|X_n| > \epsilon) \to 0 \quad$ as $n \to \infty$.
2. $X_n \rightsquigarrow Z$ means that $\mathbb{P}(X_n \leq z) \to \mathbb{P}(Z \leq z)$ at all continuity points z.
3. $X_n = O_P(a_n)$ means that, X_n/a_n is bounded in probability: for every $\epsilon > 0$ there is an $M > 0$ such that, for all large n, $\mathbb{P}\left(\left|\frac{X_n}{a_n}\right| > M\right) \leq \epsilon$.
4. $X_n = o_p(a_n)$ means that X_n/a_n goes to 0 in probability: for every $\epsilon > 0$

$$\mathbb{P}\left(\left|\frac{X_n}{a_n}\right| > \epsilon\right) \to 0 \quad \text{as } n \to \infty.$$

5. Law of large numbers: $X_1, \ldots, X_n \sim P$ then

$$\overline{X}_n \xrightarrow{P} \mu$$

   where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[X_i]$.
6. Central limit theorem: $X_1, \ldots, X_n \sim P$ then

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightsquigarrow N(0,1)$$

   where $\sigma^2 = \text{Var}(X_i)$.

2 Basic Statistics

1. Bias and Variance. Let $\widehat{\theta}$ be an estimator of θ. Then

$$\mathbb{E}(\widehat{\theta} - \theta)^2 = \text{bias}^2 + \text{Var}$$

   where $\text{bias} = \mathbb{E}[\widehat{\theta}] - \theta$ and $\text{Var} = \text{Var}(\widehat{\theta})$.
2. Maximum Likelihood. Parametric model $\{p_\theta : \theta \in \Theta\}$. We also write $p_\theta(x) = p(x;\theta)$. Let $X_1, \ldots, X_n \sim p_\theta$. MLE $\widehat{\theta}_n$ (maximum likelihood estimator) maximizes the likelihood function

$$\mathcal{L}(\theta) = \prod_{i=1}^n p(X_i;\theta).$$

3. Fisher information $I_n(\theta) = nI(\theta)$ where

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log p(X;\theta)}{\partial \theta^2}\right].$$

4. Then

$$\frac{\widehat{\theta}_n - \theta}{s_n} \rightsquigarrow N(0,1)$$

where $s_n = \sqrt{\frac{1}{nI(\widehat{\theta})}}$.

5. Asymptotic $1-\alpha$ confidence interval $C_n = \widehat{\theta}_n \pm z_{\alpha/2}\ s_n$. Then

$$\mathbb{P}(\theta \in C_n) \to 1 - \alpha.$$

6. Minimax Risk. Let $\mathcal{P}$ be a set of distributions. Let θ be a parameter and let $L(\widehat{\theta}, \theta)$ be a loss function. The minimax risk is

$$R_n = \inf_{\widehat{\theta}} \sup_{P \in \mathcal{P}} \mathbb{E}_P[L(\widehat{\theta}, \theta)].$$

If

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[L(\widehat{\theta}, \theta)] = R_n$$

then $\widehat{\theta}$ is a minimax estimator.

3 Regression

1. $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$ and prediction risk is

$$\mathbb{E}(Y - m(X))^2.$$

2. Minimizer is $m(x) = \mathbb{E}(Y|X = x)$.
3. Best linear predictor: minimize

$$\mathbb{E}(Y - \beta^T X)^2$$

where $X(1) = 1$ so that β_1 is the intercept. Minimizer is

$$\beta = \Lambda^{-1}\alpha$$

where $\Lambda(j,k) = \mathbb{E}[X(j)X(k)]$ and $\alpha(j) = \mathbb{E}(YX(j))$.

4. The data are

$$(X_1, Y_1), \ldots, (X_n, Y_n).$$

Given new X predict Y.

5. Minimize training error
$$\widehat{R}(\beta) = \frac{1}{n}\sum_i (Y_i - \beta^T X_i)^2.$$
Solution: least squares:
$$\widehat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$$
where $\mathbb{X}(i,j) = X_i(j)$.
6. Predicted values $\widehat{Y} = \mathbb{X}\widehat{\beta} = HY$ where $H = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ is the hat matrix: the projector onto the column space of $\mathbb{X}$.
7. Bias-Variance tradeoff: Write $Y = m(X) + \epsilon$ and let $\widehat{Y} = \widehat{m}(X)$ where $\widehat{m}(x) = x^T\widehat{\beta}$. Then
$$R = \mathbb{E}(\widehat{Y} - Y)^2 = \sigma^2 + \int b^2(x)p(x)dx + \int v(x)p(x)dx$$
where $b(x) = \mathbb{E}[\widehat{m}(x)] - m(x)$, $v(x) = \mathrm{Var}(\widehat{m}(x))$ and $\sigma^2 = \mathrm{Var}(\epsilon)$.

4 Classification

1. $X \in \mathbb{R}^d$ and $Y \in \{0,1\}$.
2. Classifier $h : \mathbb{R}^d \to \{0,1\}$.
3. Risk:
$$R(h) = \mathbb{P}(Y \neq h(X)).$$
Bayes rule minimizes $R(h)$:
$$h(x) = I(m(x) > 1/2) = I(\pi_1 p_1(x) > \pi_0 p_0(x))$$
where $m(x) = \mathbb{P}(Y = 1|X = x)$, $\pi_1 = \mathbb{P}(Y = 1)$, $\pi_0 = \mathbb{P}(Y = 0)$, $p_1(x) = p(x|Y = 1)$ and $p_0(x) = p(x|Y = 0)$.
4. Now code Y as $Y \in \{-1, +1\}$. Then many classifiers can be written as
$$h(x) = \mathrm{sign}(\psi(x))$$
for some ψ. For linear classifiers, $\psi(x) = \beta^T x$.
5. The risk is
$$R = \mathbb{P}(Y \neq h(X)) = \mathbb{P}(Y\psi(X) < 0)$$
which is a non-convex function of $y\psi(x)$.
6. Often we replace the loss function with a convex surrogate function such as the logistic
$$L(Y, \psi(X)) = \log(1 + \exp(-Y\psi(X))),$$
the adaboost loss
$$L(Y, \psi(X)) = \exp(-Y\psi(X))$$

or the hinge loss

$$L(Y, \psi(X)) = [1 - Y\psi(X)]_+.$$

The linear classifier that minimizes the hinge loss is called a *support vector machine (SVM)*.

5 Some Things We did Not Cover in 705

Here are some things we did not cover in 705 but you have probably come across them. We will cover them in detail in this course.

1. High dimensional linear regression. We want to predict Y using $\beta^T X$ where $X \in \mathbb{R}^d$ with $d > n$. We will consider three approaches:
   (a) Ridge regression: minimize

   $$\frac{1}{n}\sum_i (Y_i - \beta^T X_i)^2 + \lambda \sum_j \beta_j^2.$$

   Minimizer:

   $$\widehat{\beta} = (X^T X + \lambda I)^{-1} X^T Y.$$

   λ small: low bias-high variance. λ big: high bias-low variance. Choose λ by CV.
   (b) Greedy variable selection: choose $S \subset \{1, \ldots, d\}$. Bias-variance trade-off. Large S means large variance, low bias. Small S means small variance, large bias. Choosing the best subset is NP-hard so we use a greedy method (forward stepwise regression).
   (c) Lasso. Best subset selection minimizes

   $$\frac{1}{n}\sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_0.$$

   Instead of a greedy approximation, we use convex relaxation: minimize

   $$\frac{1}{n}\sum_i (Y_i - \beta^T X_i)^2 + \lambda ||\beta||_1$$

   where $||\beta||_1 = \sum_j |\beta_j|$. Returns a sparse estimator. Choose λ by CV.
2. Nonparametric regression. We want to estimate $m(x) = \mathbb{E}[Y|x = x]$ assuming that $m \in \mathcal{M}$ where $\mathcal{M}$ is a large function space. A simple estimator is the kernel regression estimator

   $$\widehat{m}(x) = \frac{\sum_i Y_i K_h(x, X_i)}{\sum_i K_h(x, X_i)}$$

   where K_h is a kernel such as $K_h(x, y) = e^{-||x-y||^2/(2h)}$. We will study this and other estimators. We also want to know the minimax risk for this problem.

3. Risk Estimation/Cross-Validation. We will have to estimate the risk to choose tuning parameters. For example: fit on part of the data. Estimate risk from held out data. Flavors: data splitting, k-fold, leave-one-out. For leave-one-out:

$$\begin{aligned}\widehat{R} &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_{(-i)})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{Y}_i}{1-H_{ii}}\right)^2 \\ &\approx \left(\frac{1}{1-\frac{d}{n}}\right)^2 \frac{1}{n}\sum_i (\widehat{Y}_i - Y_i)^2 \qquad \text{GCV} \\ &\approx \frac{1}{n}\sum_i (\widehat{Y}_i - Y_i)^2 + 2d\widehat{\sigma}^2 \qquad C_p.\end{aligned}$$

   How accurate are these methods?

6 Why Study Theory?

Inventing machine learning methods is easy. But the methods are not useful unless we understand when they work, and when they fail. Theory provides us with the ability to answer questions like the following:

1. Why is one classifier better than another?
2. Why do some prediction methods work well in certain high dimensional problems?
3. How do we chose tuning parameters in prediction algorithms?
4. Which is more important, choosing a good prediction algorithm or choosing the tuning parameters within a given algorithm?
5. Under what assumptions does a predictor work well? What is the best any method can do under these assumptions?

A good example is deep learning. Some people think this is a great breakthrough. Others think it is smoke and mirrors. Rigorous theory is still missing. Without it, all we have are examples.