36-708 Introduction and Review

1 Statistics versus ML

Statistics and ML are overlapping fields. Both address the same question: how do we extract information from data? But there are differences in emphasis. In particular, some topics get greater emphasis than others. Here are some examples:

More emphasis in ML	More emphasis in Stat	Common Areas
Bandits	Confidence Sets	Prediction (Regression and Classification)
Reinforcement Learning	Large Sample Theory	Probability Bounds (Concentration)
Efficient Computation	Statistical Optimality	Clustering
Deep Learning	Causality	Graphical Models

However, the lines between the two fields are blurry and will become increasingly so.

Another difference between the two fields is that ML researchers tend to publish short papers in conferences while Statisticians tend to publish long papers in journals. Each has advantages and disadvantages.

2 Concentration

Hoeffding's inequality:

Theorem 1 (Hoeffding) *If $Z_1, Z_2, \ldots, Z_n$ are iid with mean μ and $\mathbb{P}(a \leq Z_i \leq b) = 1$, then for any $\epsilon > 0$*

$$\mathbb{P}(|\overline{Z}_n - \mu| > \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2} \tag{1}$$

where and $\overline{Z}_n = \frac{1}{n}\sum_{i=1}^n Z_i$.

VC Dimension. Let $\mathcal{A}$ be a class of sets. If F is a finite set, let $s(\mathcal{A}, F)$ be the number of subset of F 'picked out' by $\mathcal{A}$. Define the growth function

$$s_n(\mathcal{A}) = \sup_{|F|=n} s(\mathcal{A}, F).$$

Note that $s_n(\mathcal{A}) \leq 2^n$. The *VC dimension* of a class of set $\mathcal{A}$ is

$$\mathsf{VC}(\mathcal{A}) = \sup\Big\{n : \ s_n(\mathcal{A}) = 2^n\Big\}. \tag{2}$$

If the VC dimension is finite, then there is a phase transition in the growth function from exponential to polynomial:

Theorem 2 (Sauer's Theorem) *Suppose that $\mathcal{A}$ has finite VC dimension d. Then, for all $n \geq d$,*

$$s(\mathcal{A}, n) \leq \left(\frac{en}{d}\right)^d. \tag{3}$$

Given data $Z_1, \ldots, Z_n \sim P$. The empirical measure P_n is

$$P_n(A) = \frac{1}{n}\sum_i I(Z_i \in A).$$

Theorem 3 (Vapnik and Chervonenkis) *Let $\mathcal{A}$ be a class of sets. For any $t > \sqrt{2/n}$,*

$$\mathbb{P}\left(\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| > t\right) \leq 4\, s(\mathcal{A}, 2n) e^{-nt^2/8} \tag{4}$$

and hence, with probability at least $1 - \delta$,

$$\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \leq \sqrt{\frac{8}{n} \log\left(\frac{4\, s(\mathcal{A}, 2n)}{\delta}\right)}. \tag{5}$$

Hence, if $\mathcal{A}$ has finite VC dimension d then

$$\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \leq \sqrt{\frac{8}{n}\left(\log\left(\frac{4}{\delta}\right) + d \log\left(\frac{ne}{d}\right)\right)}. \tag{6}$$

Bernstein's inequality is a more refined inequality than Hoeffding's inequality. It is especially useful when the variance of Y is small. Suppose that $Y_1, \ldots, Y_n$ are iid with mean μ, $\mathrm{Var}(Y_i) \leq \sigma^2$ and $|Y_i| \leq M$. Then

$$\mathbb{P}(|\overline{Y} - \mu| > \epsilon) \leq 2 \exp\left\{-\frac{n\epsilon^2}{2\sigma^2 + 2M\epsilon/3}\right\}. \tag{7}$$

It follows that

$$P\left(|\overline{Y} - \mu| > \frac{t}{n\epsilon} + \frac{\epsilon\sigma^2}{2(1-c)}\right) \leq e^{-t}$$

for small enough ϵ and c.

3 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 $\sigma^2 = \mathrm{Var}(X_i)$.

4 Basic Statistics

1. **Bias and Variance**. Let $\widehat{\theta}$ be an estimator of θ. Then
$$\mathbb{E}(\widehat{\theta} - \theta)^2 = \mathrm{bias}^2 + \mathrm{Var}$$
where bias $= \mathbb{E}[\widehat{\theta}] - \theta$ and Var $= \mathrm{Var}(\widehat{\theta})$. In many cases there is a **bias-variance** tradeoff. In parametric problems, we typically have that the standard deviation is $O(n^{-1/2})$ but the bias is $O(1/n)$ so the variability dominates. In nonparametric problems this is no longer true. We have to choose tuning parameters in classifiers and estimators to balance the bias and variance.
2. A set of distributions $\mathcal{P}$ is a **statistical model**. They can be small (parametric models) or large (nonparametric models).
3. **Confidence Sets.** Let $X_1, \ldots, X_n \sim P$ where $P \in \mathcal{P}$. Let $\theta = T(P)$ be some quantity of interest, Then $C_n = C(X_1, \ldots, X_n)$ is a $1 - \alpha$ confidence set if
$$\inf_{P \in \mathcal{P}} P(T(P) \in C_n) \geq 1 - \alpha.$$
4. **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
$$\mathcal{L}(\theta) = \prod_{i=1}^n p(X_i;\theta).$$

5. Fisher information $I_n(\theta) = nI(\theta)$ where
$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log p(X;\theta)}{\partial\theta^2}\right].$$
6. Then
$$\frac{\widehat{\theta}_n - \theta}{s_n} \rightsquigarrow N(0,1)$$
where $s_n = \sqrt{\frac{1}{nI(\widehat{\theta})}}$.
7. Asymptotic $1-\alpha$ confidence interval $C_n = \widehat{\theta}_n \pm z_{\alpha/2}\ s_n$. Then
$$\mathbb{P}(\theta \in C_n) \to 1-\alpha.$$

5 Minimaxity

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.

For example, if $X_1,\ldots,X_n \sim N(\theta,1)$ and $L(\widehat{\theta},\theta) = (\widehat{\theta}-\theta)^2$ then the minimax risk is $1/n$ and the minimax estimator is $\overline{X}_n$.

As another example, if $X_1,\ldots,X_n \sim p$ where $X_i \in \mathbb{R}^d$, $L(\widehat{p},p) = \int(\widehat{p}-p)^2$ and $p \in \mathcal{P}$, the set of densities with bounded second derivatives, then $R_n = (C/n)^{4/(4+d)}$. The kernel density estimator is minimax.

6 Regression

1. $Y \in \mathbb{R}$, $X \in \mathbb{R}^d$ and prediction risk is
$$\mathbb{E}(Y - m(X))^2.$$
We write $X = (X(1),\ldots,X(d))$.
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. Fitted 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 $\sigma^2 = \mathrm{Var}(\epsilon)$.

7 Classification

1. $X \in \mathbb{R}^d$ and $Y \in \{0,1\}$.
2. Classifier $h : \mathbb{R}^d \to \{0,1\}$.
3. Prediction risk:

$$R(h) = \mathbb{P}(Y \neq h(X)).$$

The **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. **Re-coded loss.** If we 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, $\psi(x) = \beta^T x$. Then the loss can be written as $I(Y \neq h(X)) = I(Y\psi(X) < 0)$ and risk is

$$R = \mathbb{P}(Y \neq h(X)) = \mathbb{P}(Y\psi(X) < 0)$$

.

5. **Linear Classifiers**. A linear classifier has the form $h_\beta(x) = I(\beta^T x > 0)$. (I am including a intercept in x. In other words $x = (1, x(2), \ldots, x(d))$.) Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$ there are several ways to estimate a linear classifier:
   (a) Empirical risk minimization (ERM): Choose $\widehat{\beta}$ to minimize

$$R_n(\beta) = \frac{1}{n}\sum_{i=1}^n I(Y_i \neq h_\beta(X_i)).$$

   (b) Logistic regression: use the model

$$P(Y = 1|X = x) = \frac{e^{\beta^T x}}{1 + e^{\beta^T x}} \equiv p(x, \beta).$$

   So $Y_i \sim$ Benoulii$(p(X_i, \beta))$. The likelihood function is

$$L(\beta) = \prod_i p(X_i, \beta)^{Y_i}(1 - p(X_i, \beta))^{1-Y_i}.$$

   The log-likelihood is strictly concave. So we have find the maximizer $\widehat{\beta}$ easily. It is easy to check that the classifier $I(p_{x,\widehat{\beta}} > 1/2)$ is linear.
   (c) SVM (support vector machine). Code Y as $+1$ or -1. We can write the classifier as $h_\beta(x) = \text{sign}(\psi_\beta(x))$ where $\psi_\beta(x) = x^T\beta$. As we said above, the loss can be written as $I(Y \neq h(X)) = I(Y\psi(X) < 0)$. Now replace the nonconvex loss $I(Y\psi(X) < 0)$ with the hinge-loss $[1 - Y_i\psi_\beta(X_i)]_+$. We minimize the regularized loss

$$\sum_{i=1}^n [1 - Y_i\psi_\beta(X_i)]_+ + \lambda||\beta||^2.$$

6. The SVM is an example of the general idea of replacing the true loss with a surrogate loss that is easier to minimize. Replacing $I(Y\psi(X) < 0)$ with

$$L(Y, \psi(X)) = \log(1 + \exp(-Y\psi(X)))$$

   gives back logistic regression. The adaboost algorithm uses

$$L(Y, \psi(X)) = \exp(-Y\psi(X)).$$

   And, as we said above, the SVM uses the hinge loss

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