CS761 Spring 2013 Advanced Machine Learning

Basics of Statistical Machine Learning

Lecturer: Xiaojin Zhu *jerryzhu@cs.wisc.edu*

Modern machine learning is rooted in statistics. You will find many familiar concepts here with a different name.

1 Parametric vs. Nonparametric Statistical Models

A *statistical model* $\mathcal{H}$ is a set of distributions.

★ *In machine learning, we call $\mathcal{H}$ the hypothesis space.*

A *parametric model* is one that can be parametrized by a finite number of parameters. We write the PDF $f(x) = f(x;\theta)$ to emphasize the parameter $\theta \in \mathbb{R}^d$. In general,

$$\mathcal{H} = \left\{ f(x;\theta) : \theta \in \Theta \subset \mathbb{R}^d \right\} \tag{1}$$

where Θ is the *parameter space*. We will often use the notation

$$\mathbb{E}_\theta(g) = \int_x g(x) f(x;\theta)\, dx \tag{2}$$

to denote the expectation of a function g with respect to $f(x;\theta)$. Note the subscript in $\mathbb{E}_\theta$ does *not* mean integrating over all θ.
★ *This notation is standard but unfortunately confusing. It means "w.r.t. this fixed θ." In machine learning terms, this means w.r.t. different training sets all sampled from this θ. We will see integration over all θ when discussing Bayesian methods.*

Example 1 *Consider the parametric model $\mathcal{H} = \{N(\mu, 1) : \mu \in \mathbb{R}\}$. Given iid data $x_1, \ldots, x_n$, the optimal estimator of the mean is $\widehat{\mu} = \frac{1}{n}\sum x_i$.*

★ *All (parametric) models are wrong. Some are more useful than others.*

A *nonparametric model* is one which cannot be parametrized by a fixed number of parameters.
Example 2 *Consider the nonparametric model $\mathcal{H} = \{P : Var_P(X) < \infty\}$. Given iid data $x_1, \ldots, x_n$, the optimal estimator of the mean is again $\widehat{\mu} = \frac{1}{n}\sum x_i$.*

Example 3 *In a naive Bayes classifier we are interested in computing the conditional $p(y|x;\theta) \propto p(y;\theta)\prod_i^d p(x_i|y;\theta)$. Is this a parametric or nonparametric model? The model is specified by $\mathcal{H} = \{p(x, y;\theta)\}$ where θ contains the parameter for the class prior multinomial distribution $p(y)$ (finite number of parameters), and the class conditional distributions $p(x_i|y)$ for each dimension. The latter can be parametric (such as a multinomial over the vocabulary, or a Gaussian), or nonparametric (such as 1D kernel density estimation). Therefore, naive Bayes can be either parametric or nonparametric, although in practice the former is more common.*

★ *Should we prefer parametric or nonparametric models? Nonparametric makes weaker model assumptions and thus is preferred. However, parametric models converges faster and are more practical.*

In machine learning we are often interested in a function of the distribution $T(F)$, for example, the mean. We call T the statistical functional, viewing F the distribution itself a function of x. However, we will also abuse the notation and say $\theta = T(F)$ is a "parameter" even for nonparametric models.

2 Estimation

Given $X_1 \dots X_n \sim F \in \mathcal{H}$, an *estimator* $\widehat{\theta}_n$ is any function of $X_1 \dots X_n$ that attempts to estimate a parameter θ.

★ *This is the "learning" in machine learning! In machine learning, a familiar case is classification where* $X_i = (x_i, y_i)$ *and* $\widehat{\theta}_n$ *is the parameters of the classifier learned from such training data. Note the subscript* n *for the training set size. Clearly,* $\widehat{\theta}_n$ *is a random variable because the training set is random.*

★ *Also note that the phrase "training set" is a misnomer because it is not a set: we allow multiple instances of the same element. Some people therefore prefer the term "training sample."*

An estimator is *consistent* if

$$\widehat{\theta}_n \xrightarrow{P} \theta. \tag{3}$$

★ *Consistency is a fundamental, desirable property of good machine learning algorithms. Here the sequence of random variables is w.r.t. training set size. Would you like a learning algorithm which gets worse with more training data?*

Because $\widehat{\theta}_n$ is a random variable, we can talk about its expectation:

$$\mathbb{E}_\theta(\widehat{\theta}_n) \tag{4}$$

where $\mathbb{E}_\theta$ is w.r.t. the joint distribution $f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$. Then, the *bias* of the estimator is

$$\text{bias}(\widehat{\theta}_n) = \mathbb{E}_\theta(\widehat{\theta}_n) - \theta. \tag{5}$$

An estimator is *unbiased* if $\text{bias}(\widehat{\theta}_n) = 0$. The *standard error* of an estimator is

$$\text{se}(\widehat{\theta}_n) = \sqrt{\text{Var}_\theta(\widehat{\theta}_n)}. \tag{6}$$

Example 4 *Don't confuse standard error with standard deviation of* f*. Let* $\hat{\mu} = \frac{1}{n}\sum_i x_i$*, where* $x_i \sim N(0,1)$*. Then the standard deviation of* x_i *is 1 regardless of* n*. In contrast,* $\text{se}(\hat{\mu}) = 1/\sqrt{n} = n^{-\frac{1}{2}}$ *which decreases with* n*.*

The *mean squared error* of an estimator is

$$\text{mse}(\widehat{\theta}_n) = \mathbb{E}_\theta\left((\widehat{\theta}_n - \theta)^2\right). \tag{7}$$

Theorem 1 $\text{mse}(\widehat{\theta}_n) = \text{bias}^2(\widehat{\theta}_n) + \text{se}^2(\widehat{\theta}_n) = \text{bias}^2(\widehat{\theta}_n) + \text{Var}_\theta(\widehat{\theta}_n)$.

★ *If* $\text{bias}(\widehat{\theta}_n) \to 0$ *and* $\text{Var}_\theta(\widehat{\theta}_n) \to 0$ *then* $\text{mse}(\widehat{\theta}_n) \to 0$*. This implied* $\widehat{\theta}_n \xrightarrow{L^2} \theta$*, and* $\widehat{\theta}_n \xrightarrow{P} \theta$*, so that* $\widehat{\theta}_n$ *is consistent.*

★ *Why are we interested in the mse? Normally we don't. We will see other "quality measures" later.*

3 Maximum Likelihood

For parametric statistical models, a common estimator is the *maximum likelihood estimator*. Let $x_1, \dots, x_n$ be iid with PDF $f(x; \theta)$ where $\theta \in \Theta$. The *likelihood function* is

$$L_n(\theta) = f(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta). \tag{8}$$

The *log likelihood function* is $\ell_n(\theta) = \log L_n(\theta)$. The maximum likelihood estimator (MLE) is

$$\widehat{\theta}_n = \mathrm{argmax}_{\theta\in\Theta} L_n(\theta) = \mathrm{argmax}_{\theta\in\Theta} \ell_n(\theta). \tag{9}$$

Example 5 *The MLE for p(head) from n coin flips is count(head)/n, sometimes called "estimating probability by the frequency." This is also true for multinomials. The MLE for $X_1, \ldots, X_N \sim N(\mu, \sigma^2)$ is $\widehat{\mu} = 1/n \sum_i X_i$ and $\widehat{\sigma}^2 = 1/n \sum (X_i - \widehat{\mu})^2$. These agree with our intuition. However, the MLE does not always agree with intuition. For example, the MLE for $X_1, \ldots, X_n \sim \mathrm{uniform}(0, \theta)$ is $\widehat{\theta} = \max(X_1, \ldots, X_n)$. You would think θ is larger, no?*

The MLE has several nice properties. The Kullback-Leibler divergence between two PDFs is

$$KL(f\|g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx. \tag{10}$$

The model $\mathcal{H}$ is *identifiable* if $\forall \theta, \psi \in \Theta$, $\theta \neq \psi$ implies $KL(f(x;\phi)\|f(x;\psi)) > 0$. That is, different parameters correspond to different PDFs.

Theorem 2 *When $\mathcal{H}$ is identifiable, under certain conditions (see Wasserman Theorem 9.13), the MLE $\widehat{\theta}_n \overset{P}{\to} \theta^*$, where θ^* is the true value of the parameter θ. That is, the MLE is consistent.*

Given n iid observations, the *Fisher information* is defined as

$$I_n(\theta) = n\mathbb{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f(X;\theta)\right)^2\right] = -n\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)\right] \tag{11}$$

Example 6 *Consider n iid observations $x_i \in \{0, 1\}$ from a Bernoulli distribution with true parameter p. $f(x;p) = p^x(1-p)^{1-x}$. It follows that $\frac{\partial^2}{\partial\theta^2}\ln f(X;\theta)$, evaluated at p, is $-x/p^2 - (1-x)/(1-p)^2$. Taking the expectation over x under $f(x;p)$ and multiply by $-n$, we arrive at $I_n(p) = \frac{n}{p(1-p)}$.*

★ *Informally, Fisher information measures the curvature of the log likelihood function around θ. A sharp peak around θ means the true parameter is distinct and should be easier to learn from n samples. The Fisher information is sometimes used in active learning to select queries. Note Fisher information is not a random variable. It does not depend on the particular n items, but rather only on the size n. Fisher information is not Shannon information.*

Theorem 3 *(Asymptotic Normality of the MLE). Let $se = \sqrt{Var_\theta(\widehat{\theta}_n)}$. Under appropriate regularity conditions, $se \approx \sqrt{1/I_n(\theta)}$, and*

$$\frac{\widehat{\theta}_n - \theta}{se} \rightsquigarrow N(0, 1). \tag{12}$$

Furthermore, let $\widehat{se} = \sqrt{1/I_n(\widehat{\theta}_n)}$. Then

$$\frac{\widehat{\theta}_n - \theta}{\widehat{se}} \rightsquigarrow N(0, 1). \tag{13}$$

★ *This says that the MLE is distributed asymptotically as $N(\theta, \frac{1}{I_n(\widehat{\theta}_n)})$. There is uncertainty, determined by both the sample size n and the Fisher information. It turns out that this uncertainty is fundamental, that no (unbiased) estimators can do better than this. This is captured by the Cramér-Rao bound. In other words, no (unbiased) machine learning algorithms can estimate the true parameter any better. Such information is very useful for designing machine learning algorithms.*

Theorem 4 *(Cramér-Rao Lower Bound) Let $\widehat{\theta}_n$ be any unbiased estimator (not necessarily the MLE) of θ. Then the variance is lower bounded by the inverse Fisher information:*

$$Var_\theta(\widehat{\theta}_n) \geq \frac{1}{I_n(\theta)}. \tag{14}$$

The Fisher information can be generalized to the high dimensional case. Let θ be a parameter vector. The Fisher information matrix has i, jth element

$$I_{ij}(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln f(X;\theta)}{\partial\theta_i \partial\theta_j}\right]. \tag{15}$$

An unbiased estimator that achieves the Cramér-Rao lower bound is said to be *efficient*. It is *asymptotically efficient* if it achieves the bound as $n \to \infty$.

Theorem 5 *The MLE is asymptotically efficient.*

★ *However, a biased estimator can sometimes achieve lower mse.*

4 Bayesian Inference

The statistical methods discussed so far are *frequentist methods*:

- Probability refers to limiting relative frequency.
- Data are random.
- Estimators are random because they are functions of data.
- Parameters are fixed, unknown constants not subject to probabilistic statements.
- Procedures are subject to probabilistic statements, for example 95% confidence intervals traps the true parameter value 95

★ *Classifiers, even learned with deterministic procedures, are random because the training set is random. PAC bound is similarly frequentist. Most procedures in machine learning are frequentist methods.*

An alternative is the *Bayesian approach*:

- Probability refers to degree of belief.
- Inference about a parameter θ is by producing a probability distributions on it. Typically, one starts with a *prior* distribution $p(\theta)$. One also chooses a *likelihood function* $p(x \mid \theta)$ – note this is a function of θ, not x. After observing data x, one applies the Bayes Theorem to obtain the *posterior* distribution $p(\theta \mid x)$:

  $$p(\theta \mid x) = \frac{p(\theta)p(x \mid \theta)}{\int p(\theta')p(x \mid \theta')d\theta'} \propto p(\theta)p(x \mid \theta), \tag{16}$$

  where $Z \equiv \int p(\theta')p(x \mid \theta')d\theta'$ is known as the *normalizing constant*. The posterior distribution is a complete characterization of the parameter.

  Sometimes, one uses the mode of the posterior as a simple point estimate, known as the *maximum a posteriori* (MAP) estimate of the parameter:

  $$\theta^{MAP} = \text{argmax}_\theta p(\theta \mid x). \tag{17}$$

  Note MAP is not a proper Bayesian approach.

- Prediction under an unknown parameter is done by integrating it out:

$$p(x \mid Data) = \int p(x \mid \theta) p(\theta \mid Data) d\theta. \tag{18}$$

★ *Here lies the major difference between frequentist and Bayesian approaches in machine learning practice. A frequentist approach would produce a point estimate $\hat{\theta}$ from Data, and predict with $p(x \mid \hat{\theta})$. In contrast, the Bayesian approach needs to integrate over different θs. In general, this integration is intractable and hence Bayesian machine learning has been focused on either finding special distributions for which the integration is tractable, or finding efficient approximations.*

Example 7 *Let θ be a d-dim multinomial parameter. Let the prior be a Dirichlet $p(\theta) = Dir(\alpha_1, \ldots, \alpha_d)$. The likelihood is multinomial $p(x \mid \theta) = Multi(x \mid \theta)$, where x is a "training" count vector. These two distributions are called conjugate to each other as the posterior is again Dirichlet: $p(\theta \mid x) = Dir(\alpha_1 + x_1, \ldots, \alpha_d + x_d)$.*

Now let's look into the predictive distribution for some "test" count vector x'. If $\theta \sim Dir(\beta)$, the result of integrating θ out is

$$\begin{aligned} p(x' \mid \beta) &= \int p(x' \mid \theta) p(\theta \mid \beta) d\theta & (19) \\ &= \frac{(\sum_k x'_k)!}{\prod_k (x'_k!)} \frac{\Gamma\left(\sum_k \beta_k\right)}{\Gamma\left(\sum_k \beta_k + x'_k\right)} \prod_k \frac{\Gamma\left(\beta_k + x'_k\right)}{\Gamma\left(\beta_k\right)} & (20) \end{aligned}$$

This is an example where the integration has a happy ending: it has a simple(?) closed-form. This is known as a Dirichlet compound multinomial distribution, also known as a multivariate Pólya distribution.

Where does the prior $p(\theta)$ come from?

- Ideally it comes from domain knowledge. One major advantage of Bayesian approaches is the principled way to incorporate prior knowledge in the form of the prior.

- Non-informative, or flat, prior, where there does not seem to be a reason to prefer any particular parameter. This may however create *improper priors*. Let $X \sim N(\theta, \sigma^2)$ with σ^2 known. A flat prior $p(\theta) \propto c > 0$ would be improper because $\int p(\theta) d\theta = \infty$, so it is not a density. Nonetheless, the posterior distribution is well-defined.

  A flat prior is not transformation invariant. Jeffrey's prior $p(\theta) \propto I(\theta)^{1/2}$ is.

- It should be pointed out that in practice, the choice of prior is often dictated by computational convenience, in particular conjugacy.