36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 34: December 2

Lecturer: Siva Balakrishnan

34.1 Review and Outline

In the last class we were discussing classification:

1. Generative and discriminative classifiers
2. Naive Bayes
3. Empirical Risk Minimization

Today we will discuss (briefly) the problem of sampling and integration. This is Chapter 24 of the Wasserman book.

34.2 Motivation

The methods we discuss today are particularly useful in Bayesian inference. The basic idea in Bayesian inference is that we are doing parameter estimation, and we have a prior $\pi(\theta)$ and we observe data $(X_1, \ldots, X_n)$. We then compute the posterior distribution:

$$\pi(\theta|X_1, \ldots, X_n) = \frac{\mathcal{L}(\theta|X_1, \ldots, X_n)\pi(\theta)}{\int_\theta \mathcal{L}(\theta|X_1, \ldots, X_n)\pi(\theta)d\theta},$$

and use it in various ways (point estimation - take posterior mean, "confidence" intervals - use posterior intervals).

Suppose we consider point estimation: we want to compute the mean of the posterior:

$$\widehat{\theta} = \int \theta \ \pi(\theta|X_1, \ldots, X_n)d\theta.$$

The main difficulty here is that the normalizing constant:

$$\int_\theta \mathcal{L}(\theta|X_1, \ldots, X_n)\pi(\theta)d\theta$$

is difficult to compute so we do not really "know" the posterior.

One simple idea is that if we could sample $(\theta_1, \ldots, \theta_m)$ from the posterior then we could try to approximate:

$$\widehat{\theta} \approx \frac{1}{m} \sum_{i=1}^{m} \theta_i.$$

Of course, sampling from the posterior could be as hard as computing the distribution. Markov Chain Monte Carlo is a way to sample from the posterior without computing the normalizing constant.

34.3 Monte Carlo Integration

You have explored this idea previously in an assignment. Suppose I want to (approximately) compute an expectation:

$$\mu = \mathbb{E}_{X \sim P}[f(X)],$$

but cannot do so analytically. However, suppose that I can sample $X_1, \ldots, X_n \sim P$, then I could approximate this expectation as:

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

This is the idea of Monte Carlo Integration.

A first question is how accurate is this approximation? We know by the CLT that $\widehat{\mu} - \mu$ will be close to normally distributed so it only remains to calculate the variance.

$$\mathrm{Var}(\widehat{\mu}) = \frac{1}{n} \mathrm{Var}(f(X)),$$

which is for instance small if f is bounded, i.e if $|f| \leq M$, in which case:

$$\mathrm{Var}(f(X)) \leq M^2,$$

and the variance is roughly $1/n$. So if we take enough samples, then our estimate will be quite good. Furthermore, we can estimate this variance in the usual way:

$$\widehat{\mathrm{Var}}(\widehat{\mu}) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \widehat{\mu})^2.$$

Example 1: Suppose we want to compute the standard Gaussian CDF $\Phi(x)$ at some point x. This is given by:

$$\mu = \Phi(x) = \int_{-\infty}^{x} f(u)du,$$

where $f(x)$ is the standard Gaussian pdf. We can re-write this as an expectation:

$$\mu = \int_{u} h(u)f(u)du,$$

where $h(x) = \mathbb{I}(u \leq x)$. In order to use Monte Carlo integration, we would just draw many samples from a standard Gaussian, and then use:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x).$$

34.4 Importance Weighting

A next scenario is one where we cannot directly sample from the distribution P but still want to approximate:

$$\mu = \mathbb{E}_{X\sim P}[f(X)].$$

Let us suppose that we can instead sample from some distribution Q. Suppose further that we can evaluate ratios of the form:

$$P(X)/Q(X),$$

for any value X. We can see that:

$$\mu = \mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{X\sim Q}\left[\frac{P(X)}{Q(X)}f(X)\right] = \mathbb{E}_{X\sim Q}[w(X)f(X)],$$

at least provided that the weights are never infinite. This means that we can use Monte Carlo integration, given samples $X_1, \ldots, X_n$,

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} w(X_i)f(X_i).$$

The variance of this estimate depends on both the variance of w and the variance of f.

An important point is that even when we can sample from P it might be desirable to use importance sampling, as it might have much smaller variance. Let us consider a simple example:

Example 2: Suppose we want to estimate

$$\mu = \mathbb{P}(Z > 3) = 0.0013,$$

where $Z \sim N(0,1)$. We can write this as an expectation as before, and then use Monte Carlo. Using $n = 100$, we find (from simulating many times) that $\mathbb{E}(\widehat{\mu}) = .0015$ and $\mathrm{Var}(\widehat{\mu}) = .0039$. An important observation is that most of the samples are wasted since very few samples are in the right tail (i.e. bigger than 3).

Suppose we instead used importance sampling, and sampled from Q which is $N(4,1)$ (a shifted normal). In this case we find that $\mathbb{E}(\widehat{\mu}) = .0011$ and $\mathrm{Var}(\widehat{\mu}) = .0002$. Importance sampling reduces the variance by a factor of 20.

The rough guideline is that the variance of importance sampling is smallest if we sample proportional to $p|f|$, i.e. we sample more where both p and $|f|$ are large.

34.5 Markov Chain Monte Carlo

In the Bayesian problem we began with we cannot really use either Monte Carlo or importance weighting. We instead use MCMC.

First, the big picture: In Markov Chain Monte Carlo (MCMC) the goal is to design a Markov Chain, whose stationary distribution is the distribution P under which we want to compute an integral. We then sample from the Markov Chain, and then use the law of large numbers (adapted to Markov Chains) to argue that our estimate is good.

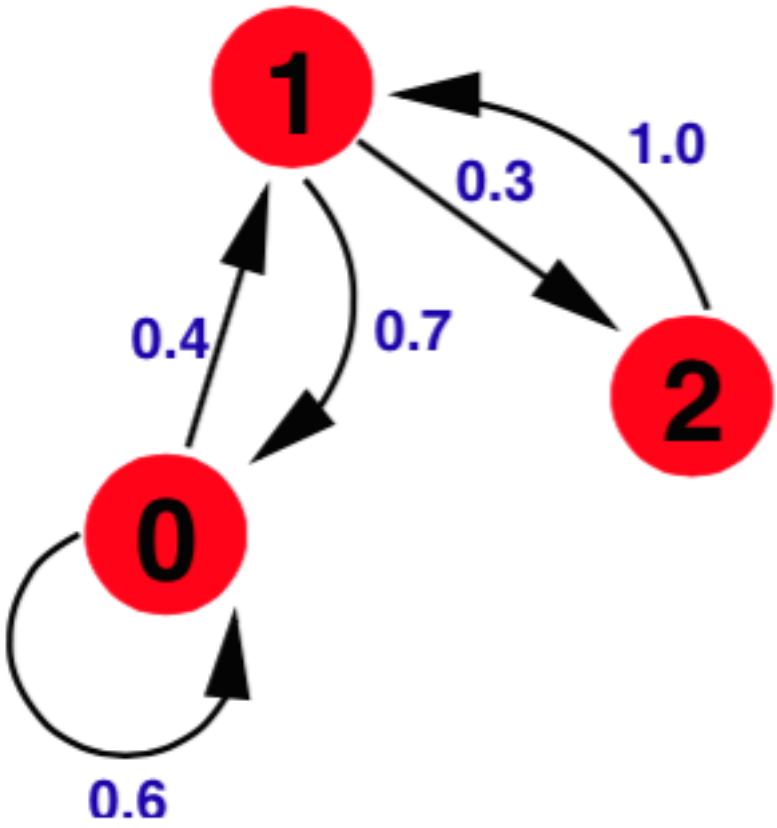
The details are a bit involved so we will get started today and finish up in the next lecture. We will do this at a very high-level.

First, what is a Markov Chain? A Markov Chain is a collection of random variables $\{X_1, \ldots, X_n\}$ that forms a graphical model: $X_1 \to X_2 \to \ldots \to X_n$.

In order to specify the joint distribution of a Markov Chain, we need to specify $p(X_1)$ and $p(X_{i+1}|X_i)$. To do this conveniently the focus is usually on what are called time-homogenous Markov Chains, i.e. $p(X_{i+1}|X_i) = T(X_i, X_{i+1})$, for a function T that does not depend on i. This function T is called the transition matrix or transition kernel of the Markov Chain.

Let us consider a simple example: suppose all the variables are discrete and take values $\{0, 1, 2\}$.

Consider a diagram of the form:



This diagram is specifying the transition matrix for us. Particularly, it says that the probability: $P(X_{i+1} = 2|X_i = 1) = 0.3$.

The next thing, we need to know is that Markov Chains almost always have a stationary distribution. The stationary distribution roughly, is the distribution of the random variable X_n for n large. We denote it by π, i.e.

$$\pi(x) = P(\lim_{n\to\infty} X_n = x).$$

An important aspect of Markov Chains is that they forget their initial state (i.e. they are ergodic/they mix), so π does not depend on $p(X_1)$.

Markov Chains also have a LLN associated with them: If $\{X_1, \ldots, X_n\}$ is a Markov Chain with stationary distribution π, then

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \int_x \pi(x)f(x)dx = \mathbb{E}_\pi[f(x)].$$

This is a pretty magical property. The Markov Chain samples are actually dependent, i.e. X_2 depends on X_1, and X_3 depends on X_2 and X_1, and so on. The LLN says that even though there are these dependencies, they are weak (and get much weaker as you get further

away in the chain), so samples from a Markov Chain behave quite similarly to i.i.d. samples from the stationary distribution.

With all of these preliminaries in place only one thing remains: construct a Markov Chain with a particular stationary distribution (think, the posterior distribution). We will consider this problem in the next lecture.