NYU Lecture 2/12/19

Bayesian Statistics

Outline

1. Terms and Background
2. Some useful distributions
3. Bayes' Theorem
4. Priors and Posteriors (+ example)
5. MLE, MAP, and duality
6. Posterior predictive

1. Terms

Average \( \overline{\theta} = \frac{\theta_1 + \theta_2 + \ldots + \theta_n}{n} \)

Logarithm \( \log(x) = y \iff e^y = x \)
- Only valid for \( x > 0 \)
- Monotone: \( x > x' \iff \log(x) > \log(x') \)

Distribution
- Roughly a function from some number(s) to a nonnegative number.
- Integrates to 1 over its support
- Integral: summation over continuous space
  - Ex: \( \int_0^1 x \, dx \) "sums" \( x \) values from 0 to 1, each with weight \( dx \).
- Support: All valid input values to a function.
  - Ex: Support of \( \log \) is positive reals
Generic Forms of Distributions

- **Marginal:** \( P(Y) \) \( \rightarrow P_\theta(Y) = P(Y; \theta) \)
- **Conditional:** \( P(Y \mid X) \) (may have parameters) \( \rightarrow P_\theta(Y \mid X) = P(Y \mid X; \theta) \)
- **Joint:** \( P(Y, X) \) \( \rightarrow P_\theta(Y, X) = P(Y, X; \theta) \)

Properties

- **Unordered Joints** \( P(Y, X) = P(X, Y) \)
- **Chain Rule** \( P(Y, X) = P(Y \mid X) P(X) \)
- **Marginalization** \( P(Y) = \int P(Y \mid X) P(X) \, dx = \sum_x P(Y, X) \) \( x \leftarrow \text{support of } X \text{ is some set } \mathcal{X} \)

- \( X \) has been "Marginalized out"
- Weighted sum where \( P(X) \) now determines the weights

Maximum Likelihood Estimation (MLE)

\[ \hat{\theta} = \text{Maximize } P_\theta(X) \]

In words: We specify a distribution \( P(X) \) that tells us how likely observing a value of \( X \) would be. But \( P(X) \) actually takes some parameters \( \theta \), which we do not know. We can find the most likely value of \( \theta \) by choosing the one that maximizes the likelihood of \( X \) (where \( X \) is our data).
Properties of MLE
- Constants with respect to $\Theta$ do not change $\hat{\Theta}$
- Ex: $\hat{\Theta} = \max_\Theta P_\Theta(x) = \max_\Theta c P_\Theta(x) + d$ $\uparrow$ $c > 0$
- Could also be functions of $x$ that do not depend on $\Theta$
  $\hat{\Theta} = \max_\Theta P_\Theta(x) = \max_\Theta f(x) P_\Theta(x) + g(x)$ $\uparrow$
  $f(x) > 0$ for all $x$.
- Monotone transformations do not change $\hat{\Theta}$
  $\hat{\Theta} = \max_\Theta P_\Theta(x) = \max_\Theta \log(P_\Theta(x))$

2) Useful distributions

We've already seen one example of a distribution, the Normal (AKA Gaussian) distribution:

$$\mathcal{N}(X; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi} \sigma} \exp\left(-\frac{1}{2\sigma^2} (X - \mu)^2\right)$$

In our generic notation, $\Theta = (\mu, \sigma^2)$ and $P_\Theta = \mathcal{N}(\mu, \sigma^2)$.

Normal distribution properties
- Real valued support, symmetric
- Saw this distribution used in Linear Regression

What if your observations have different support?
$X$ is Binary $\{0, 1\} \Rightarrow $ Bernoulli $(x; \theta) = \theta^x (1-\theta)^{1-x}$
- $\theta$ is the mean probability that $X = 1$.

$X$ is a count $\{0, 1, 2, \ldots\} \Rightarrow $ Poisson $(x; \lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$
- $\lambda$ is the mean of the distribution over counts.

Both Poisson and Bernoulli deal with discrete variables, other distributions deal with continuous variables:

$X$ is between 0 and 1 $[0, 1] \Rightarrow $ Beta $(x; a, b) = \frac{x^{a-1} (1-x)^{b-1}}{B(a, b)}$
- $B$ is the "Beta Function" \( B(a, b) \) normalizes the numerator so it integrates to 1.

$X$ is positive real $\mathbb{R}^+ \Rightarrow $ Gamma $(x; \alpha, \beta) = \frac{\beta^x}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$
- $\Gamma$ is the "Gamma Function", also a normalizer.
Bayes Theorem

Suppose I gather some data

$$X_1, \ldots, X_n \quad (n = 10 \text{ samples})$$

and I assume $X_i$ comes from a normal distribution. Further, assume I know $\Sigma$ so we only want to estimate the mean $\mu$.

MLE approach gives us $\hat{\mu} = \bar{X}$ (the average of the data).

$$\hat{\mu} = \max_{\mu} \frac{1}{n} \sum_{i=1}^{n} N(x_i; \mu, \sigma^2) = \max_{\mu} \frac{1}{n} \sum_{i=1}^{n} \log(N(x_i; \mu, \sigma^2))$$

**WRITE OUT**

$$\text{Normal} = \max_{\mu} \frac{1}{n} \sum_{i=1}^{n} \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right) \right)$$

$$\log(AB) = \log(A) + \log(B)$$

$$\log(\frac{A}{B}) = \log(A) - \log(B)$$

Left term is a constant (W.R.T. $\mu$)

$$\max_{\mu} \frac{1}{n} \sum_{i=1}^{n} \log \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( -\frac{1}{2\sigma^2} (x_i - \mu)^2 \right) \right)$$

Log ($\exp(A)$) = $A$

$$\max_{\mu} \frac{1}{n} \sum_{i=1}^{n} -\frac{1}{2\sigma^2} (x_i - \mu)^2$$

$\frac{1}{2\sigma^2}$ is a constant

$$\max_{\mu} \frac{1}{n} \sum_{i=1}^{n} -(x_i - \mu)^2$$

max $f(x) = \min_{\mu} \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$

Take derivative and set equal to zero to solve

$$\nabla \mu = -2 \sum_{i=1}^{n} (x_i - \mu) = 0$$
Divide both sides by -2: 
\[ \sum_{i=1}^{n} (x_i - \mu) = 0 \]

Expand summation: 
\[ \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \mu = 0 \]

\[ \sum_{i=1}^{n} x_i = n \mu \]

Divide both sides by n: 
\[ \frac{1}{n} \sum_{i=1}^{n} x_i - \mu = 0 \]

Add \( \mu \) to both sides: 
\[ \frac{1}{n} \sum_{i=1}^{n} x_i = \mu \]

Left hand side is the def. of average: 
\[ \bar{x} = \mu. \]

Suppose someone now comes along and tells you \( \mu \) lies between 2 and 4 with high certainty.

The Bayesian guest is to use the additional info, which is not precise, combine it with the data info, and to come up with a different (better) estimator.

The key is to turn expert statements into one of probability.

Roughly \( P(2 < \mu < 4) \approx 1 \).
There are many ways to encode the info. about $\mu$. One possibility is to take the distribution for $\mu$ to be a normal distribution with mean $\nu = 3$ and standard deviation $\tau = \frac{1}{2}$.

$\Rightarrow N(\mu; \nu, \tau^2)$

$\Rightarrow \frac{\tau}{2} \tau$ covers about 95% of the distribution so $P(2 < \mu < 4) \approx 0.95$, which is close to our goal.

$\Rightarrow$ still allows a small chance that the expert is wrong.

How do we include this information?

Bayes Theorem.

Consider the joint distribution over $X$, the data, and $\mu$, the unknown parameter: $P(X, \mu)$

Recall the chain rule:

$P(X, \mu) = P(X|\mu) P(\mu) = P(\mu | X) P(X)$  [by the unordered property]

$\Rightarrow P(\mu | X) = \frac{P(X|\mu) P(\mu)}{P(X)} \leq$ Bayes Theorem.

More generally:

$P(\text{Parameters} | \text{Data}) = \frac{P(\text{Data} | \text{Parameters}) P(\text{Parameters})}{P(\text{Data})}$
Prior $P(\text{Parameters})$: Prior beliefs about the parameters, before seeing any data.

- In our example, our prior was $\mathcal{N}(\mu; \nu, \sigma^2)$.

Likelihood $P(\text{data} | \text{Parameters})$: If I know the parameters, how likely is the data?

- In our example, our likelihood was $\prod_{i=1}^{n} \mathcal{N}(x_i; \mu, \sigma^2)$.

Posterior $P(\text{Parameters} | \text{data})$: Beliefs about the parameters after seeing the data.

- This is what we care about.
- In the example, we observed $x_1, \ldots, x_{10}$ and now we want to combine our prior beliefs about $\mu$ with our data to get a new, updated set of beliefs.

Marginal Likelihood $P(\text{data})$: Normalizing constant

- Independent of $\mu$
- If our goal is to find the most likely $\mu$, we get to ignore this.
Example Revisited

\[ X_1, \ldots, X_{10} \quad (n=10) \]

\[ P(X_i | \mu) = N(X_i; \mu, \tau^2) \]

\[ P(\mu) = N(\mu; \nu, \tau^2) \]

"Posterior Inference" is about taking your data and priors and finding the posterior:

\[ P(\mu | X_1, \ldots, X_{10}) = \frac{P(X_1, \ldots, X_{10} | \mu)P(\mu)}{P(X_1, \ldots, X_{10})} \]

\[ \frac{\prod_{i=1}^{10} N(X_i; \mu, \tau^2)}{P(X_1, \ldots, X_{10})} \]

\[ N(\mu; \nu, \tau^2) \]

We know \( P(X_1, \ldots, X_{10}) \) won't change no matter the value of \( \mu \) because it doesn't take \( \mu \) as an input. So let's consider it a constant

\[ Z = P(X_1, \ldots, X_{10}) \]

For now we'll work with just the numerator, and put a pin here that our answer is proportional to the posterior, but the true posterior needs \( Z \) to ensure it integrates to 1.

\[ P(\mu | X_1, \ldots, X_{10}) \propto \left[ \prod_{i=1}^{10} N(X_i; \mu, \tau^2) \right] N(\mu; \nu, \tau^2) \]

If we encounter other constants along the way to simplifying the above, we can drop those too and assume they're part of \( Z \).
\[ P(\mu | x_1, \ldots, x_n) \propto \left[ \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi} \sigma_i} \exp\left( -\frac{1}{2\sigma_i^2} (x_i - \mu)^2 \right) \right] N(\mu; \nu, \tau^2) \]

Drop constants w.r.t. \( \mu \) and \( \frac{1}{\sqrt{2\pi} \sigma_i} \)

\[ = \left[ \prod_{i=1}^{n} \exp\left( -\frac{1}{2\sigma_i^2} (x_i - \mu)^2 \right) \right] \left[ \exp\left( -\frac{1}{2\tau^2} (\mu - \nu)^2 \right) \right] \]

\[ \exp(\mathbf{A}) \exp(\mathbf{B}) = \exp(\mathbf{A} + \mathbf{B}) \]

Expand quadratics \((\mathbf{A} - \mathbf{B})^2 = \mathbf{A}^2 - 2\mathbf{AB} + \mathbf{B}^2\)

\[ \alpha \exp\left( \sum_{i=1}^{n} \frac{1}{2\sigma_i^2} (x_i^2 - 2x_i \mu + \mu^2) - \frac{1}{2\tau^2} (\mu - \nu)^2 \right) \]

\[ \cong \exp\left( \sum_{i=1}^{n} \frac{1}{2\sigma_i^2} (-2x_i \mu + \mu^2) - \frac{1}{2\tau^2} (\mu^2 - 2\nu \mu) \right) \]

\[ \exp(\mathbf{A} + \text{const}) = \exp(\mathbf{A}) \exp(\text{const}) \]

Expand out the summation and move \(-\frac{1}{2\sigma_i^2}\) outside

\[ \sum_{i=1}^{n} x_i = n \bar{x} \quad \text{and} \quad \sum_{i=1}^{n} \mu_i^2 = n \mu^2 \]

\[ \alpha \exp\left( -\frac{1}{2\sigma_i^2} (-2n \bar{x} \mu + n \mu^2) - \frac{1}{2\tau^2} (\mu^2 - 2\nu \mu) \right) \]

Factor out \(-\frac{1}{2}\)

\[ \alpha \exp\left( -\frac{1}{2} \left[ \frac{1}{\sigma_i^2} (-2n \bar{x} \mu + n \mu^2) + \frac{1}{\tau^2} (\mu^2 - 2\nu \mu) \right] \right) \]
Group together $-2\mu$ terms and $\mu^2$ terms

Factor out $\frac{1}{\sqrt{2\pi n + \frac{1}{x^2}}}\exp\left(-\frac{1}{2} \left[ \frac{1}{\sqrt{2\pi n + \frac{1}{x^2}}} \left( -2\mu \left( \frac{1}{\sqrt{2\pi n + \frac{1}{x^2}}} \right) + \mu^2 \right) \right]ight)$

Replace with $\bar{v} = \frac{\frac{1}{\sqrt{2\pi n + \frac{1}{x^2}}} \left( -2\mu \right) + \mu^2}{\sqrt{2\pi n + \frac{1}{x^2}}}$

$\bar{v}$ is a const. w.r.t. $\mu$, so

$\exp(\mu \bar{v} + \mu^2) \propto \exp(\mu \bar{v} + \mu^2) \exp(\bar{v}^2)$

$\propto \exp(\mu \bar{v} + \mu^2 + \bar{v}^2)$

Factor the quadratic $A^2 + B^2 - 2AB = (A - B)^2$

$\propto \exp\left(-\frac{1}{2} \left[ \frac{1}{\sqrt{2\pi n + \frac{1}{x^2}}} \right] \left( \mu - \bar{v} \right)^2 \right)$

$\propto \exp\left(-\frac{1}{2\bar{v}^2} \left( \mu - \bar{v} \right)^2 \right)$

$\propto N(\mu; \bar{v}, \frac{1}{\bar{v}^2})$  

up to a constant, this is just a normal distribution!

Note: Since this is a valid distribution, $\bar{v} = 1$
What Just happened?
- We started with a belief about $\mu$ that it was normally distributed with mean $\bar{V}$ and variance $\sigma^2$.
- We observed samples $x_1, \ldots, x_{10}$ from a likelihood also normally distributed, but with mean $\mu$ and variance $\sigma^2$.
- The posterior for $\mu$ was normal and centered at mean $\bar{V}$:

$$
\bar{V} = \frac{\frac{1}{\sigma^2} \sum x_i + \frac{1}{\tau^2} V}{\frac{1}{\sigma^2 n} + \frac{1}{\tau^2}}
$$

This is called "conjugacy": your prior and posterior are in the same class of distributions (i.e. a Normal distribution here).
- Doesn't always happen with any prior and any likelihood.
- When it does, you get a nice, analytic solution like the above one.
What else happened?

- The MLE is $\bar{x} = \bar{\mu}$, just the mean of the data.
- The most likely posterior value is the peak of $N(\mu; \tilde{\nu}, \frac{\sigma^2}{\nu})$, which is just $\tilde{\nu}$:
  $$\tilde{\mu} = \max_{\mu} N(\mu; \tilde{\nu}, \frac{\sigma^2}{\nu}) = \tilde{\nu}$$
- This is called the “maximum a posteriori” (MAP) estimate. It’s different than the mean.
- Let’s unpack it:
  $$\tilde{\nu} = \frac{\frac{1}{\sigma^2} \bar{x} + \frac{1}{\nu} \nu}{\frac{1}{\sigma^2} + \frac{1}{\nu}}$$
- Weighted combination of the mean $\bar{x}$ of the data and the mean $\nu$ from the prior.
- The weights come from two sources:
  a) the inverse of the variances, $\frac{1}{\sigma^2}$ and $\frac{1}{\nu}$
  b) the number of observations, $n$
- We can think of the prior variance as quantifying our uncertainty about $\mu$ before seeing the data, so if we have high uncertainty, $\sigma^2$ is big, $\frac{1}{\sigma^2}$ is small, and the prior mean gets little weight.
The more certainty we have on our prior, the more samples we need to move toward the data \( X \).

This is called "Shrinkage".

We are shrinking the mean towards a prior belief.

\[
\begin{align*}
\text{Ex. Data (x)} & & \text{Prior (\mu)} \\
\bar{x} = 2.75 & & \mu \text{ very far from } \bar{x} \\
\text{Posterior} & & \text{so } 2.75 \text{ seems very unlikely}
\end{align*}
\]

\[\hat{\mu} = 3.15\]

Duality

The steps we took to get our MAP estimate \( \hat{\mu} \) were:

1. Specify prior and likelihood
2. Calculate posterior
3. Find Max. \( P(\mu | x) \) [the \( \mu \) that has highest posterior probability].

But this is a bit indirect.
Direct MAP estimation:
1. Specify prior and likelihood
2. \( \tilde{\mu} = \max_{\mu} \left[ \prod_{i=1}^{n} N(x_i; \mu, \sigma^2) \right] N(\mu; \nu, \tau^2) \)

If we do the same algebra tricks as before (take the log, drop constants, negate + min., etc), we end up with:

\[
\min_{\mu} \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{\tau^2} (\mu - \nu)^2
\]

Divide by \( \frac{1}{\sigma^2} \): \[
\min_{\mu} \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{\sigma^2}{\tau^2} (\mu - \nu)^2
\]

Let \( \lambda = \frac{\sigma^2}{\tau^2} \):
\[
\min_{\mu} \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + \lambda (\mu - \nu)^2
\]

If we had a prior that \( \mu \) was mean zero, then \( \nu = 0 \):
\[
\min_{\mu} \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + 2\lambda \mu^2
\]

\( \lambda \) is a regularizer, so it turns out there is an interesting deep connection between regularized MLE and MAP estimation. They're sort of the same thing:
- Likelihoods correspond to loss functions
- Priors correspond to regularizers
- Penalty weight (\( \lambda \)) corresponds to how uncertain you are in your prior beliefs (inverse prior variance).

When you choose to regularize, think about the prior beliefs you are implicitly making!