Probabilistic Artificial Intelligence

A. Krause

Lecture 1 - Probability Review

1.1 Probability Space

Okay so We start with defining a **probability space**, and the idea with that thing is just to give You everything You need to do probabalistic stuff. Without further ado, a probability space is a triple consisting of $(\Omega, \mathcal{A}, P)$.

Ω is called the **sample space** and is just a set of all possible outcomes of an experiment. For the flip of two coins, You get $\Omega = \{HH, HT, TH, TT\}$.

$\mathcal{A}$ is the **event space**, and it is usually thought of as every possible combination of elements of Ω. But why is it defined this way? Well, the idea is that *combinations* of outcomes are events. For instance, an event might be getting one or more tail, and that event corresponds to a subset $\{TH, HT, TT\} \in \Omega$. You can come up with much more complicated events in case of more complicated sample spaces. Point here is that We just have a set of all possible events, which We will now use in:

P - think of P as function $P : \mathcal{A} \to [0, 1]$. All P does is assign a probability for each event in $\mathcal{A}$.

With those three constructs We can talk about outcomes of an experiment, group outcomes of an experiment into events, and talk about the probabilities of those events. There are of course some other constraints for this all to make sense, like $P(\Omega) = 1$ and $\forall S \subset \Omega : P(S) \in [0, 1]$).

Random variables then can be thought of as functions - let X be a random variable, then $X : \mathcal{A} \to \mathbb{R}$, roughly. The output range depends on the variable - is it discrete, continuous, is it univariate or multivariate etc.

Then You have three axioms for it all to work:

Normalization: $P(\Omega) = 1$, so probability of all possible outcomes must equal to 1.

Non-negativity: $S \in \mathcal{A} \implies P(S) \geq 0$.

σ-additivity - probabilities of disjoint events can simply be added.

Then You can come up with stories for random variables and get distributions like Bernoulli for 1 flip of a coin, Binomial for many flips, multinomial for dice outcomes etc.

If the variable is continuous You get a PDF and a CDF, PDF must be non-negative and CDF must integrate to 1 etc.

Then You get Your Gaussian, Your vector of random variables and a **joint distribution** (meaning You specify a value for each variable in vector).

Then You get Your conditional distribution defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

In it's simplest form. That becomes Bayes' Rule if You observe the **product rule** $P(B|A)P(A) = P(A \cap B)$. You can keep taking stuff out too, so

$$P(A, B, C) = P(A|B, C)P(B|C)P(C)$$

That's the chain product rule.

The other important rule is the **sum rule** (or marginalization), in that

$$P(X_1 = y, X_2 = z) = \sum_{x \in \sup X_3} P(X_3 = x)P(X_1 = y)P(X_2 = z)$$

While We're here, let's also clear up **prior, likelihood** and **posterior** probabilities.

I suppose talking about distributions makes sense - the prior distribution of X is, well, the distribution of X without any additional information. Just what We know about X before the experiment.

Likelihood is $P(E|X = x)$ and should really be thought of as a function that takes the value of X, i.e. what X turned out to be, and it gives You the probability of evidence E given that $X = x$. In English one might say that $P(E|X = x)$ is the likelihood that We observe E given $X = x$. X is usually some hidden variable(s) that We say generated evidence E.

Posterior is the same as conditional probability - $P(X|E)$, or distribution of X given that some evidence is true. The distinction between posterior and likelihood is that in likelihood We are talking about the likelihood of something else given the variable We care about X, where as in the posterior We are interested in adjusting the distribution of our variable X given some evidence.

Then there's independence and conditional independence, no worries there.

High dimensional (binary in the most trivial case) multivariate distributions require an exponential number of parameters to fully specify. Marginalizing out variables runs into the same problem, so, We have as problems:

Representation (how to represent high dimensional distributions in a not parameter-exponential way), learning (given data, learn the distribution that produced it) and inference (given a distribution, make predictions)

1.2 Gaussians

$$P(X = x) = \frac{1}{2\pi\sqrt{|\Sigma|}} \cdot \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

Lecture 2 - Bayesian Linear Regression

The idea is simple - ordinary linear regression (and variants thereof) yield a point estimate $\hat{y}$. We'd like to know how uncertain We are about this estimate.

2.1 Ridge Regression as Bayesian Inferece

Recall that We can find the coefficients for the best linear fit by doing

$$X^\top(X\mathbf{w} - \mathbf{y}) = 0 \tag{1}$$
$$X^\top X\mathbf{w} - X^\top\mathbf{y} = 0 \tag{2}$$
$$X^\top X\mathbf{w} = X^\top\mathbf{y} \tag{3}$$
$$\mathbf{w} = (X^\top X)^{-1}X^\top\mathbf{y} \tag{4}$$

Is the way that usually goes. Now, in order to get the solution for ridge regression We do

$$\mathbf{w} = (X^\top X + \lambda I)^{-1}X^\top\mathbf{y} \tag{5}$$

Is the modification for ridge regression to penalize large coefficients in $\mathbf{w}$.

So, now, let's make things probabilistic:

For $\mathbf{w}$, let's assume that $\mathbf{w} \sim \mathcal{N}(0, \sigma^2_{\mathbf{W}})$ and also let's assume that $\mathbf{w} \perp \mathbf{x}_i \; \forall i \in [n]$, so We have a normal prior over the weights and a-priori the weights are independent of the data - without knowing anything about $\mathbf{x}_i$, this is all We've got.

Then let's make the standard

$$P(\mathbf{y}_i|\mathbf{w}, \mathbf{x}_i) \sim \mathcal{N}(\mathbf{w}^\top\mathbf{x}_i, \sigma^2_{\mathbf{y}}) \tag{6}$$

So We're assuming that given the one true weight vector, our labels are normally distributed around the predicted mean. They are also of course independent, and all have the same variance.

So now the idea is - We have a prior on the weights, and We have a likelihood function which involves the weights. By the glory of Bayes' rule, that's enough to try to calculate a posterior on the weights:

$$P(\mathbf{w}|\mathbf{x}_{1\ldots n}, \mathbf{y}_{1\ldots n}) = \frac{1}{z} \cdot P(\mathbf{w}, \mathbf{x}_{1\ldots n}, \mathbf{y}_{1\ldots n}) \tag{7}$$
$$= \frac{1}{z} \cdot P(\mathbf{x}_{1\ldots n}) \cdot p(\mathbf{w}|\mathbf{x}_{1\ldots n}) \cdot P(\mathbf{y}_{1\ldots n}|\mathbf{w}, \mathbf{x}_{1\ldots n}) \tag{8}$$
$$= \frac{1}{z'} \cdot p(\mathbf{w}|\mathbf{x}_{1\ldots n}) \cdot P(\mathbf{y}_{1\ldots n}|\mathbf{w}, \mathbf{x}_{1\ldots n}) \tag{9}$$
$$= \frac{1}{z'}\mathcal{N}(0, \sigma^2_{\mathbf{W}}) \cdot \prod_{i=1}^{n} \mathcal{N}(\mathbf{w}^\top\mathbf{x}_i, \sigma^2_{\mathbf{y}}) \tag{10}$$
$$= \frac{1}{z'}\frac{1}{z_{\mathbf{w}}} \exp\left(-\frac{1}{\sigma^2_{\mathbf{w}}}\|\mathbf{w}\|^2\right) \cdot \frac{1}{z_{\mathbf{y}}} \prod_{i=1}^{n} \exp\left(-\frac{1}{\sigma^2_{\mathbf{y}}} \cdot \|y_i - \mathbf{w}^\top\mathbf{x}_i\|^2\right) \tag{11}$$

So, at 7 We are simply using the whole $P(A|B) = P(A,B)/P(B)$ thing, to rearrange the terms any way We like - the point is to have one conjunction above and one below, and factorize the above conjunction with the chain rule in a way that can leverage our assumptions.

Then at 9, We absorb that $P(\mathbf{x}_{1\ldots n})$ term since it's kind of irrelevant - just some Gaussian that We'll ultimately not care about.

Finally We use an assumption or two.

Now, note that when fitting $\mathbf{w}$, We're going to maximize our result. Since We just take about an extrema, We can start stripping parts away:

$$\arg\max_{\mathbf{w}} = \frac{1}{z'}\frac{1}{z_{\mathbf{w}}}\exp\left(-\frac{1}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2\right)\cdot\frac{1}{z_{\mathbf{y}}}\prod_{i=1}^{n}\exp\left(-\frac{1}{\sigma_{\mathbf{y}}^2}\cdot\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2\right) \tag{12}$$

$$= \exp\left(-\frac{1}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2\right)\cdot\prod_{i=1}^{n}\exp\left(-\frac{1}{\sigma_{\mathbf{y}}^2}\cdot\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2\right) \tag{13}$$

$$= \exp\left(-\frac{1}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2\right)\cdot\prod_{i=1}^{n}\exp\left(-\frac{1}{\sigma_{\mathbf{y}}^2}\cdot\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2\right) \tag{14}$$

$$= \exp\left(-\frac{1}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2-\sum_{i=1}^{n}\frac{1}{\sigma_{\mathbf{y}}^2}\cdot\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2\right) \tag{15}$$

$$= -\frac{1}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2-\sum_{i=1}^{n}\frac{1}{\sigma_{\mathbf{y}}^2}\cdot\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2 \tag{16}$$

$$\arg\min_{\mathbf{w}} = \frac{\sigma_{\mathbf{y}}^2}{\sigma_{\mathbf{w}}^2}\|\mathbf{w}\|^2+\sum_{i=1}^{n}\|y_i-\mathbf{w}^\top\mathbf{x}_i\|^2 \tag{17}$$

Where in 16, We just multiplied by the positive constant $\sigma_{\mathbf{y}}^2$ to get a clean coefficient for the λ term.

Anyway this shows that if We choose $\lambda = \frac{\sigma_{\mathbf{y}}^2}{\sigma_{\mathbf{w}}^2}$, then the solution to the ridge regression problem is equivalent to finding the maximum a posteriori solution. ($P(\mathbf{w}|\mathbf{x}_{1\ldots n}, \mathbf{y}_{1\ldots n})$ is the posterior in question).

2.2 Distribution of the weights

Okay, so, We have that under Bayesian regression

$$\mathbf{w} = \mathbf{w} = (X^\top X + \lambda I)^{-1}X^\top\mathbf{y} \tag{18}$$

$$= \mathbf{w} = (X^\top X + \frac{\sigma_{\mathbf{y}}^2}{\sigma_{\mathbf{w}}^2}I)^{-1}X^\top\mathbf{y} \tag{19}$$

$$\tag{20}$$