EECS E6720 Bayesian Models for Machine Learning

Columbia University, Fall 2016

Lecture 2, 9/15/2016

Instructor: John Paisley

- Next, we look at another instance of a conjugate prior. To help motivate the practical usefulness of the distributions we will consider, we discuss this prior in the context of a regression problem.

Problem setup

- We're often given a data set of the form $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. The goal is to learn a prediction rule from this data so that, given a new $\hat{x}$ we can predict its associated an unobserved $\hat{y}$. This is called a regression problem.

- We often refer to x as "covariates" or "features" and y as a "response."

Linear regression

- One model for this problem is to assume a linear relationship between the inputs x and outputs y such that
$$y_i = x_i^T w + \epsilon_i$$
The term ϵ_i accounts for the fact that we usually can't find a w such that $y_i = x_i^T w$ for all i. The model value of interest is w and we simply set $\epsilon_i = y_i - x_i^T w$. Further assumptions let $w \in \mathbb{R}^d$ and $\epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$.

Likelihood term for y

- Using only this information, we have an implied likelihood term of y given X and w (where $X = \{x_i\}$).

- First, notice that for each i
$$y_i \overset{ind}{\sim} \text{Normal}(x_i^T w, \sigma^2)$$
where we recall that $\text{Normal}(y_i | x_i^T w, \sigma^2) = (2\pi\sigma^2)^{\frac{1}{2}} \exp\{-\frac{1}{2\sigma^2}(y_i - x_i^T w)^2\}$.

- Because of the conditional independence assumption of y_i given w, we can write the likelihood as

$$p(y_1, \ldots, y_N|w, X) = \prod_{i=1}^{N} p(y_i|w, x_i)$$

- The vector w is the model parameter that is unknown as we want to learn it. Bayesian linear regression takes the additional step of treating w as a random variable with a prior distribution. Hence in the Bayesian setting we refer to w as a model variable instead of a model parameter.

- After defining a prior $p(w)$, we use Bayes rule to learn a posterior distribution on w:

$$p(w|X, \vec{y}) = \frac{p(\vec{y}|X, w)p(w)}{\int_{\mathbb{R}^d} p(\vec{y}|X, w)p(w)dw} = \frac{\prod_{i=1}^{N} p(y_i|x_i, w)p(w)}{\int_{\mathbb{R}^d} \prod_{i=1}^{N} p(y_i|x_i, w)p(w)dw}$$

- Question: Why is X always being conditioned on? (i.e., on the RHS of |)

- Answer: Look at what we have distributions on.
  - The model assumes a generative distribution for y
  - We put a prior distribution on w
  - We haven't assumed any distribution on x_i
  - In short, this results from the *model assumptions* that we have made

<u>Prior on w</u>

- We still have to define the prior distribution on w. *Out of convenience* we define

$$w \sim \text{Normal}(0, \lambda^{-1}I)$$

Why do we do this?
Because it's conjugate to the likelihood. As a result, the posterior distribution of w is a recognizable distribution with known parameters that can be easily calculated from the data.

<u>Posterior of w</u>

- We next calculate the posterior distribution of w for this model. The steps we take are as follows:

Using Bayes rule we have that

$$p(w|X, \vec{y}) \propto \prod_{i=1}^{N} \text{Normal}(y_i|x_i^T w, \sigma^2)\text{Normal}(w|0, \lambda^{-1}I)$$

By direction plugging in using the form of a Gaussian, we have that

$$p(w|X, \vec{y}) \propto \left[\prod_{i=1}^{N} \mathrm{e}^{-\frac{1}{2\sigma^2}(y_i - x_i^T w)^2}\right]\left[\mathrm{e}^{-\frac{\lambda}{2}w^T w}\right]$$

We don't need to include the constants out front $(2\pi\sigma^2)^{-\frac{1}{2}}$ and $(\lambda/(2\pi))^{\frac{1}{2}}$ because they don't depend on the unknown part of the distribution, w. When writing out the normalizing constant as the integral of the numerator, one can see that these terms appear in the numerator and denominator and cancel out. The only reason we do this is for convenience (it's less to write).

I'll work through the full derivation below. You can skip to Step 7 for the final result.

1. The first step to calculate $p(w|X,\vec{y})$ is to combine into one exponential by summing the terms in the exponent,

$$p(w|X,\vec{y}) \propto \mathrm{e}^{-\frac{\lambda}{2}w^Tw-\frac{1}{2\sigma^2}\sum_{i=1}^N(y_i-x_i^Tw)^2}$$

2. In the next step, we expand the square to write an equation quadratic in w

$$p(w|X,\vec{y}) \propto \exp\Big\{-\tfrac{1}{2}\Big[w^T(\lambda I+\tfrac{1}{\sigma^2}\textstyle\sum_i x_ix_i^T)w-2w^T(\tfrac{1}{\sigma^2}\sum_i y_ix_i)+\tfrac{1}{\sigma^2}\sum_i y_i^2\Big]\Big\}$$

3. Next, we write

$$\begin{aligned} p(w|X,\vec{y}) &\propto \exp\Big\{-\tfrac{1}{2}\Big[w^T(\lambda I+\tfrac{1}{\sigma^2}\textstyle\sum_i x_ix_i^T)w-2w^T(\tfrac{1}{\sigma^2}\sum_i y_ix_i)\Big]\Big\}\exp\Big\{-\tfrac{1}{2}\Big[\tfrac{1}{\sigma^2}\textstyle\sum_i y_i^2\Big]\Big\} \\ &\propto \exp\Big\{-\tfrac{1}{2}\Big[w^T(\lambda I+\tfrac{1}{\sigma^2}\textstyle\sum_i x_ix_i^T)w-2w^T(\tfrac{1}{\sigma^2}\sum_i y_ix_i)\Big]\Big\} \end{aligned}$$

In the first line, we just moved things around, but didn't change the actual function. We did this to justify the second line, which again takes advantage of the fact that we only care about proportionality. If "$\propto$" were instead "$=$," this would of course be mathematically wrong. However, "$\propto$" allows us to treat $\exp\{-\frac{1}{2}[\sigma^{-2}\sum_i y_i^2]\}$ as a pre-multiplying constant w.r.t. w that will cancel out when we divide the numerator by its integral over w to get the posterior distribution and return to "$=$." We get rid of it just for convenience.

4. In the fourth step, we invert this process by multiplying by a term of our choosing that's constant with respect to w. The goal is to complete the square in the exponential term. We choose to multiply this term by the somewhat arbitrary looking term

$$\exp\Big\{-\tfrac{1}{2}\Big[(\tfrac{1}{\sigma^2}\textstyle\sum_i y_ix_i)^T(\lambda I+\tfrac{1}{\sigma^2}\sum_i x_ix_i^T)^{-1}(\tfrac{1}{\sigma^2}\sum_i y_ix_i)\Big]\Big\}$$

However, notice that this doesn't involve w, so by multiplying with it we aren't violating the proportionality w.r.t. w. Think of this as working with a function of w, and scaling that function up and down until we reach the point that it integrates to 1.

5. By multiplying the last term in Step 3 with the term in Step 4 and shuffling things around, we get a long term that I will break down as follows

$$p(w|X,\vec{y}) \propto \exp\Big\{-\tfrac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\Big\}$$

where

$$\Sigma = (\lambda I+\tfrac{1}{\sigma^2}\textstyle\sum_i x_ix_i^T)^{-1} \quad \text{and} \quad \mu = \Sigma(\tfrac{1}{\sigma^2}\sum_i y_ix_i)$$

You can verify this by plugging in for μ and Σ, expanding the quadratic term, pulling out the constant and seeing that the result is the multiplication of Step 3 with Step 4.

6. Finally, we want to calculate $p(w|X,\vec{y})$ exactly. We're hoping we can find a closed form solution to the integral in

$$p(w|X,\vec{y}) = \frac{\exp\big\{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\big\}}{\int \exp\big\{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\big\}dw}$$

   where I use the same definition of μ and Σ as in Step 5. Approaching this problem from a purely mathematical standpoint, this is not easy at all. However, mathematicians and statisticians have been collecting known probability distributions for our reference. At some point, someone actually solved this integral and now we have the solution without having to recalculate the integral each time.

   Specifically, we know that a d-dimensional multivariate Gaussian distribution with mean μ and covariance Σ has the form

$$p(w|\mu,\Sigma) = (2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)}$$

   We know from other people's calculations that this function of w integrates to 1. This function is also proportional to the function in Step 5. In Step 6 we're solving the integral in order to find the constant to multiply Step 5 with to make the function integrate to 1. Therefore, we know this constant is $(2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}$ and

$$\int \exp\big\{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\big\}dw = (2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}$$

   It's an argument based on logic and based on trust that previous proofs that the nonnegative function $(2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\mathrm{e}^{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)}$ integrates to 1 are correct, which is the stopping point for this class (and 99.99% of all "Bayesians" including myself).

7. Therefore, we have shown that the posterior distribution $p(w|X,\vec{y})$ is a multivariate Gaussian with mean μ and covariance Σ, where

$$\Sigma = (\lambda I + \tfrac{1}{\sigma^2}\textstyle\sum_i x_i x_i^T)^{-1} \qquad \mu = \Sigma(\tfrac{1}{\sigma^2}\textstyle\sum_i y_i x_i)$$

   Notice that this is in the same family as the prior $p(w)$, which was also a multivariate Gaussian. Therefore we chose a *conjugate prior* for w.

Making predictions

- For many applications, the posterior distribution of a model's variables gives useful information. For example, it often provides information on the structure underlying the data, or cause and effect type relationships.

- For example, in the linear regression model $y = x^Tw + \epsilon$ we've been discussing, the vector w tells us the relationship between the inputs of x and output y. e.g., if w_k is positive the model will lead us to believe that increasing the kth dimension of x will increase y and by how much. We might then use this information to figure out which dimensions we need to "improve."

- Often, however, we just want to make predictions on new data using the model of the old data.

Predictive distribution

- For Bayesian linear regression, we ultimately want to predict a new $\hat{y}$ given its associated $\hat{x}$ and all previously observed (x, y) pairs. In other words, we wish to form the predictive distribution

$$p(\hat{y}|\hat{x}, \vec{y}, X) = \int_{\mathbb{R}^d} p(\hat{y}|\hat{x}, w)p(w|\vec{y}, X)dw$$

- Notice that the predictive distribution can be written as a marginal distribution over the model variables.

- However, implied in the left hand side is the model definition, which tells us which variables to integrate over on the right hand side, as well as the specific distributions to plug in.

- That is, just writing a predictive distribution $p(\hat{y}|\hat{x}, \vec{y}, X)$ is not enough information to know how to represent it as a marginal distribution, where marginalization (i.e., integration) takes place over the model variables. It is necessary to know in advance what the underlying model is in order to know what model variables to integrate over.

- Let's step back and think more generally about what we're doing when we construct a predictive probability distribution.

  Bayesian model: The Bayesian modeling problem is summarized in the following sequence.

$$\begin{aligned} \text{Model of data:} \quad & X \sim p(X|\theta) \\ \text{Model prior:} \quad & \theta \sim p(\theta) \\ \text{Model posterior:} \quad & p(\theta|X) = p(X|\theta)p(\theta)/p(X) \end{aligned}$$

  Predictions: The goal of making predictions is to use past data to predict the future under the same modeling assumption. This is why we choose to model data in the first place: We assume there is an underlying rule governing the way data is created whether its seen or unseen data. By defining a model we define a rule. Seeing some data lets us get a better sense of what that rule should be, letting us learn from the past. Therefore, when making predictions we assume future data follows the same rule as past data,

$$\text{Future data:} \quad \hat{x} \sim p(x|\theta) \qquad \leftarrow \text{the same model as the past data}$$

  We don't know θ, but the past data X, along with our prior believe about θ, gives us information about it. In fact we have a posterior distribution on it $p(\theta|X)$. We can use this to "score" or "weight" each possible setting of θ. This is the marginalization procedure,

$$\text{Predictive distribution:} \quad p(\hat{x}|X) = \int p(\hat{x}|\theta)p(\theta|X)d\theta$$

  So by marginalizing, we are considering the infinite number of possible settings for θ in the likelihood and weighting it by our posterior belief of how probable that setting is. For certain probability distributions, we can solve this integral analytically.

- The Bayesian linear regression model we've been discussing is one example where we can solve for the predictive distribution. Our goal is to solve the integral

$$p(\hat{y}|\hat{x}, \vec{y}, X) = \int \underbrace{(2\pi\sigma^2)^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}(\hat{y}-\hat{x}^T w)^2}}_{= p(\hat{y}|\hat{x},w)} \underbrace{(2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)}}_{= p(w|\vec{y},X)} dw$$

  Notice that we keep all the terms this time because we are writing an equality here, not a proportionality. The specific values of μ and Σ for the regression problem were derived above.

- Below I'll derive this predictive distribution. The trick this time will be to multiply and divide by the same specifically chosen value. You can skip to Step 5 for the result.

  1. First, we expand both squares and pull everything not depending on w out in front of the integral

  $$\begin{aligned} p(\hat{y}|\hat{x}, \vec{y}, X) &= (2\pi\sigma^2)^{-\frac{1}{2}} (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\hat{y}^2 - \frac{1}{2}\mu^T\Sigma^{-1}\mu} \\ &\quad \times \int \mathrm{e}^{-\frac{1}{2}\left(w^T(\hat{x}\hat{x}^T/\sigma^2 + \Sigma^{-1})w - 2w^T(\hat{x}\hat{y}/\sigma^2 + \Sigma^{-1}\mu)\right)} dw \end{aligned}$$

  2. Next we multiply and divide by a constant w.r.t. w. We want to choose a constant to result in the integral equaling 1. We notice that we can complete the square in the exponent, allowing us to put it into the form of a Gaussian distribution. Therefore, multiply and divide by the term

  $$(2\pi)^{-\frac{d}{2}} \left|\sigma^{-2}\hat{x}\hat{x}^T + \Sigma^{-1}\right|^{\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(\hat{x}\hat{y}/\sigma^2 + \Sigma^{-1}\mu)^T(\hat{x}\hat{x}^T/\sigma^2 + \Sigma^{-1})^{-1}(\hat{x}\hat{y}/\sigma^2 + \Sigma^{-1}\mu)}$$

  3. If we put this extra term in the numerator to the right of the integral and the term in the denominator to the left of the integral, we can complete the square in the right-most exponent and see that a Gaussian results. Therefore we know the integral equals 1 and we can simply remove this term. The result is

  $$p(\hat{y}|\hat{x}, \vec{y}, X) = \frac{(2\pi\sigma^2)^{-\frac{1}{2}} (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\hat{y}^2 - \frac{1}{2}\mu^T\Sigma^{-1}\mu}}{(2\pi)^{-\frac{d}{2}} \left|\sigma^{-2}\hat{x}\hat{x}^T + \Sigma^{-1}\right|^{\frac{1}{2}} \mathrm{e}^{-\frac{1}{2}(\hat{x}\hat{y}/\sigma^2 + \Sigma^{-1}\mu)^T(\hat{x}\hat{x}^T/\sigma^2 + \Sigma^{-1})^{-1}(\hat{x}\hat{y}/\sigma^2 + \Sigma^{-1}\mu)}}$$

  4. The final step uses two important matrix equalities that are often useful. In the context of the equation above, these are

  $$\left|\sigma^{-2}\hat{x}\hat{x}^T + \Sigma^{-1}\right| = |\Sigma^{-1}|(1 + \hat{x}^T\Sigma\hat{x}/\sigma^2) = |\Sigma|^{-1}(1 + \hat{x}^T\Sigma\hat{x}/\sigma^2)$$

  plugging this in above, several terms cancel out and we end up multiplying the ratio of exponents by $(2\pi)^{-\frac{1}{2}}(\sigma^2 + \hat{x}^T\Sigma\hat{x})^{-\frac{1}{2}}$ The second useful equality is called the matrix inversion lemma. This allows us to say that

  $$(\hat{x}\hat{x}^T/\sigma^2 + \Sigma^{-1})^{-1} = \Sigma + \Sigma\hat{x}(\sigma^2 + \hat{x}^T\Sigma\hat{x})^{-1}\hat{x}^T\Sigma$$

  Notice that the inverted term is a scalar, not a matrix. We plug this in where it appears in the exponent, and combine into one exponential. The bookkeeping is fairly involved, but we get terms that cancel and can write the resulting quadratic term as a square. As a result, the exponential term is $\mathrm{e}^{-\frac{1}{2(\sigma^2 + \hat{x}^T\Sigma\hat{x})}(\hat{y} - \hat{x}^T\mu)^2}$.

5. As a result, we have calculated that

$$p(\hat{y}|\hat{x}, \vec{y}, X) = (2\pi)^{-\frac{1}{2}}(\sigma^2 + \hat{x}^T\Sigma\hat{x})^{-\frac{1}{2}}e^{-\frac{1}{2(\sigma^2+\hat{x}^T\Sigma\hat{x})}(\hat{y}-\hat{x}^T\mu)^2}$$

   Notice that this is a univariate Normal distribution with mean $\hat{x}^T\mu$ and variance $\sigma^2 + \hat{x}^T\Sigma\hat{x}$. Again, the terms μ and Σ are calculated using the prior parameters and the data $\vec{y}$ and X.

Another modeling example

- Let's look at another Bayesian modeling example that is slightly more complicated.

Problem setup

- Again we're given a data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ where $x \in \Omega_x$, which is an arbitrary space, and this time $y \in \{0, 1\}$. The goal is to *classify* a new x by assigning a label to it. That is, we want to predict the unknown value of $y \in \{0, 1\}$ associated with a new input x.

Bayesian classifier model

- This approach to classification assumes a generative process for *both* y and x. For the model, we will assume that each (x, y) pair is generated i.i.d. as follows

$$\text{Model:} \quad y \overset{iid}{\sim} \text{Bernoulli}(\pi), \quad x|y \overset{ind}{\sim} p(x|\theta_y)$$

  To draw x, we need to know y, which is what $x|y$ is meant to convey.

- As is evident, the joint distribution on (x, y) is not a distribution that can be factorized into separate distributions on x and y. Rather, it must be factorized as

$$p(x, y|\Theta, \pi) = p(x|y, \Theta)p(y|\pi)$$

  This factorization says that we can first sample a value for y without knowing x, after which we sample x from a distribution that depends on the value of y and some parameters Θ. This is also what the generative process above says. In fact, there we further assume $\Theta = \{\theta_0, \theta_1\}$ and let y pick out the correct parameters from this set.

- We can think of this as follows, imagine x contains the text of an email and y indicates whether it is spam or not. The generative model assumes that spam and non-spam emails are generated from the same distribution family, but with different parameters. All spam emails share one parameter, θ_1, and all non-spam emails share another parameter θ_0. The data is generated by first choosing whether a spam or non-spam email will be generated by flipping a coin with bias π, and then generating the email itself from a distribution using the class-specific parameter.

- It should be clear that x doesn't have to be an email, but something else, for example several statistics from a patient during a health checkup that can be used to diagnose an illness. The x for this problem will be in a different domain than for the email problem, which is why it is beneficial to keep the distribution and space of x general.

Prior distributions

- The model variables are the class probability $\pi \in (0,1)$ and class-specific variables θ_1, θ_0. We next choose prior distributions for these. We select for π out of convenience and write a generic prior for each θ

$$\pi \sim \text{Beta}(a,b), \qquad \theta_0, \theta_1 \overset{iid}{\sim} p(\theta)$$

  We are making one assumption here about θ, that the class-specific variables are drawn *independently* from the same prior distribution. This will let us simplify the joint likelihood of the model for posterior computation.

Posterior computation

- Again we want to find the posterior, but this time of π, θ_0 and θ_1. Using Bayes rule,

$$p(\pi, \theta_1, \theta_0 | \vec{y}, X) = \frac{p(X, \vec{y}|\pi, \theta_1, \theta_0)p(\pi, \theta_1, \theta_0)}{\int_{\Omega_\theta}\int_{\Omega_\theta}\int_0^1 p(X, \vec{y}|\pi, \theta_1, \theta_0)p(\pi, \theta_1, \theta_0)d\pi d\theta_1 d\theta_0}$$

- This time we have three variables instead of one. However, looking closer we will see that these distributions factorize nicely. (This is the exception, not the rule.)

- Prior: The prior can be written $p(\pi, \theta_1, \theta_0) = p(\pi)p(\theta_1)p(\theta_0)$. Of course $p(\theta_1)$ and $p(\theta_0)$ are the same exact distribution, but evaluated at different points.

- Likelihood: By the modeling assumption, we can write the likelihood as

$$\begin{aligned} p(X, \vec{y}|\pi, \theta_1, \theta_0) &= \prod_{i=1}^N p(x_i, y_i|\pi, \theta_1, \theta_0) \\ &= \prod_{i=1}^N p(x_i|y_i, \pi, \theta_1, \theta_0)p(y_i|\pi, \theta_1, \theta_0) \\ &= \prod_{i=1}^N p(x_i|\theta_{y_i})p(y_i|\pi) \end{aligned}$$

  The product is from the independence assumption. The transition from the first to second line is a rule of probability that's always true. The transition from the second to third line simplifies the to reflect the dependence structure assumed by the model.

- Finally, we can return to Bayes rule and try computing it. In the numerator, we group the multiplications across the three variables

$$p(\pi, \theta_1, \theta_0 | \vec{y}, X) \propto \left[\prod_{i:y_i=1} p(x_i|\theta_1)p(\theta_1)\right]\left[\prod_{i:y_i=0} p(x_i|\theta_0)p(\theta_0)\right]\left[\prod_{i=1}^N p(y_i|\pi)p(\pi)\right]$$

- The normalizing constant is the integral of this term. Notice that since we can write this as the product of three separate functions of over each model variable, the posterior can be written as

$$p(\pi, \theta_1, \theta_0|\vec{y}, X) = \frac{\prod_{i:y_i=1} p(x_i|\theta_1)p(\theta_1)}{\int \prod_{i:y_i=1} p(x_i|\theta_1)p(\theta_1)d\theta_1} \cdot \frac{\prod_{i:y_i=0} p(x_i|\theta_0)p(\theta_0)}{\int \prod_{i:y_i=0} p(x_i|\theta_0)p(\theta_0)d\theta_0} \cdot \frac{\prod_{i=1}^{N} p(y_i|\pi)p(\pi)}{\int \prod_{i=1}^{N} p(y_i|\pi)p(\pi)d\pi}$$

  However, this is just the product of three instances of Bayes rule and is equivalently saying

$$p(\pi, \theta_1, \theta_0|\vec{y}, X) = p(\theta_1|\{x_i : y_i = 1\})p(\theta_0|\{x_i : y_i = 0\})p(\pi|\vec{y})$$

- Last week we worked through $p(\pi|\vec{y})$. We saw that this was the posterior of a beta-Bernoulli process,

$$p(\pi|\vec{y}) \propto \prod_{i=1}^{N} p(y_i|\pi)p(\pi) \quad \longrightarrow \quad p(\pi|\vec{y}) = \text{Beta}(a + \textstyle\sum_i y_i, b + N - \sum_i y_i)$$

  That is, for the class probability, we simply sum the number of times our data was in each class and use those values as the parameters in the beta distribution.

- The other class-conditional posterior distributions on θ_0 and θ_1 are problem-specific. We notice the feature of these distributions is that we partition the data set into two groups, those belonging to class 1 and those in class 0. Then, we wish to solve (for, e.g., class 1)

$$p(\theta_1|\{x_i : y_i = 1\}) \propto \prod_{i:y_i=1} p(x_i|\theta_1)p(\theta_1)$$

  In many cases we can do this. For example, when $x \in \mathbb{R}^d$ and $p(x|\theta)$ is a Gaussian, we can pick a conjugate prior for the mean and covariance.

Naive Bayes classifier

- When x is a complex multidimensional object, a simplifying assumption about $p(x|\theta)$ is sometimes made on how this can be further factorized. To give an overview, let x be composed of m pieces of information possibly in different domains $x = \{x^{(1)}, \ldots, x^{(m)}\}$. For example, some $x^{(j)}$ could real numbers and some non-negative integers. A naive Bayes classification model makes the assumption that the variables θ can be separate into m groups as well and that

$$p(x|\theta) = \prod_{j=1}^{m} p(x^{(j)}|\theta^{(j)})$$

- Separate priors are selected for each $\theta^{(j)}$, possibly in different distribution families, and we assume $\theta_0^{(j)}, \theta_1^{(0)} \overset{iid}{\sim} p_j(\theta^{(j)})$.

- The Bayes classifier is then computed as above, after which one finds (for class 1, for example)

$$p(\theta_1|\{x_i : y_i = 1\}) = \prod_{j=1}^{m} p_j(\theta_1^{(j)}|\{x_i^{(j)} : y_i = 1\})$$

- Spam detection: To make this more concrete, consider the problem of spam detection. In the spam detector model we will consider here, a pair (x, y) consists of a label $y \in \{0, 1\}$ indicating "spam" or "not spam" and x is a v dimensional vector of word counts for a vocabulary of size v. Thus $x(j)$ is a count of how many times word j (e.g., j maps to "deal") appears in the email. We need to define a distribution on x as its model, which can also be thought of as a joint probability distribution on all the values in x,

  $$p(x|\theta) \quad \Leftrightarrow \quad p(x(1), \dots, x(v)|\theta)$$

  Naive Bayes takes the additional modeling step of breaking the correlation structure between these values

  $$p(x(1), \dots, x(v)|\theta) = \prod_{j=1}^{v} p(x(j)|\theta(j)) \qquad \leftarrow \text{a modeling assumption – we define this}$$

  From a generative perspective, we are saying that for an observation x_i, each "piece" of x_i is generated independently from its own distribution given its class. The assumption that $x_i(j)$ and $x_i(j')$ are independent given its class variable θ_{y_i} is what is "naive" about this model.

  We use Bayes rule to calculate the posterior of π exactly as before. Using it to calculate the posterior of each θ_y can be split into the process of calculating the posterior of each $\theta_y(j)$ separately. For example, in spam detection we can define $p(x(j)|\theta_y(j))$ to be a Poisson distribution, and $p(\theta_y(j))$ to be a gamma distribution. These two distributions are conjugate and so the posterior of $\theta_y(j)$ will also be gamma. Because of independence assumptions, solving for the posterior of θ_y can be separated into v independent applications of Bayes rule, one for each $\theta_y(j)$.

Predictive distribution

- As with Bayesian linear regression, a major application of the Bayes classifier is to predicting the labels of new data. That is, given a *training* set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ of labeled data we want to learn a model so that we can predict the unobserved label $\hat{y} \in \{0, 1\}$ of a newly observed $\hat{x}$.
- Again we want to form the predictive distribution $p(\hat{y}|\hat{x}, X, \vec{y})$. This can be done, but is not as straightforward as before. We'll try two ways. The first way will lead to a dead-end. The second approach will work.

First attempt

- We want to calculate

  $$p(\hat{y}|\hat{x}, X, \vec{y}) = \int_{\Omega_\theta} \int_{\Omega_\theta} \int_0^1 p(\hat{y}|\hat{x}, \theta_1, \theta_0, \pi) p(\theta_1, \theta_0, \pi | X, \vec{y}) d\pi d\theta_1 d\theta_0$$

  There are two distributions we need to know. The second one is the posterior that we previously calculated and showed for this model can be factorized as

  $$p(\theta_1, \theta_0, \pi | X, \vec{y}) = p(\theta_1|X, \vec{y}) p(\theta_0|X, \vec{y}) p(\pi|\vec{y})$$

  The first term is trickier and we can't simply follow the procedure for the Bayesian linear regression model. The reason is that we are asked for a distribution on $\hat{y}$ conditioned on both π and $\hat{x}$ plus the class-specific variables. However, according to our model we have

  $$y \sim \text{Bernoulli}(\pi), \qquad x|y \sim p(x|\theta_y)$$

The generative model for the data gives us a marginal distribution on y and a conditional distribution on x. Therefore, unlike before we can't just go to the model to find out what to plug in for $p(\hat{y}|\hat{x}, \theta_1, \theta_0, \pi)$. However, we notice that these distributions in combination with Bayes rule provide us with all the information we need.

- The distribution $p(\hat{y}|\hat{x}, \theta_1, \theta_0, \pi)$ can be interpreted as the *posterior* distribution of $\hat{y}$ given $\hat{x}$ and the model variables. By Bayes rule,

$$p(\hat{y}|\hat{x}, \theta_1, \theta_0, \pi) = \frac{p(\hat{x}|\theta_{\hat{y}})p(\hat{y}|\pi)}{p(\hat{x}|\theta_1)p(\hat{y}=1|\pi) + p(\hat{x}|\theta_0)p(\hat{y}=0|\pi)}$$

Notice that because $\hat{y}$ is a discrete variable, the integral turns into a sum over the set of possible values $\hat{y}$ can take.

- We can swap $p(\hat{y}|\hat{x}, \theta_1, \theta_0, \pi)$ with the Bayes rule version of it in the integral to form the predictive distribution. We're left with

$$p(\hat{y}|\hat{x}, X, \vec{y}) = \int_{\Omega_\theta}\int_{\Omega_\theta}\int_0^1 \frac{p(\hat{x}|\theta_{\hat{y}})p(\hat{y}|\pi)p(\theta_1|X,\vec{y})p(\theta_0|X,\vec{y})p(\pi|\vec{y})}{p(\hat{x}|\theta_1)p(\hat{y}=1|\pi) + p(\hat{x}|\theta_0)p(\hat{y}=0|\pi)} d\pi d\theta_1 d\theta_0$$

However, we now hit a dead end. Even though we have actual functions we can plug in everywhere, the problem now is that there is a sum in the denominator. For example, if we plug in the functions involving π we can see that the integral is not tractable for this variable and so we can't get an analytic function for the predictive distribution.

Second attempt

- Even though the first attempt didn't work, we still have other options. We first notice that $\hat{y} \in \{0, 1\}$, meaning we can only query the predictive distribution at these two values. Above we used Bayes rule to make progress toward the final calculation. Now we use it again, but on the marginal distribution itself:

$$\begin{aligned} p(\hat{y}|\hat{x}, X, \vec{y}) &= \frac{p(\hat{x}|\hat{y}, X, \vec{y})p(\hat{y}|X, \vec{y})}{p(\hat{x}|\hat{y}=1, X, \vec{y})p(\hat{y}=1|X,\vec{y}) + p(\hat{x}|\hat{y}=0, X, \vec{y})p(\hat{y}=0|X,\vec{y})} \\ &= \frac{p(\hat{x}|\hat{y}, X, \vec{y})p(\hat{y}|\vec{y})}{p(\hat{x}|\hat{y}=1, X, \vec{y})p(\hat{y}=1|\vec{y}) + p(\hat{x}|\hat{y}=0, X, \vec{y})p(\hat{y}=0|\vec{y})} \end{aligned}$$

- Again we need to know what to plug in for the likelihood $p(\hat{x}|\hat{y}, X, \vec{y})$ and prior $p(\hat{y}|\vec{y})$. The first thing we should think is that these are both marginal distributions and can be represented as an integral over model variables.

- Likelihood: First,

$$\begin{aligned} p(\hat{x}|\hat{y}, X, \vec{y}) &= \int_{\Omega_\theta}\int_{\Omega_\theta} p(\hat{x}|\hat{y}, \theta_1, \theta_0)p(\theta_1, \theta_0|X, \vec{y})d\theta_1 d\theta_0 \\ &= \int_{\Omega_\theta} p(\hat{x}|\theta_{\hat{y}})p(\theta_{\hat{y}}|\{x_i : y_i = \hat{y}\})d\theta_{\hat{y}} \end{aligned}$$

Notice that π isn't involved because, according to the model, conditioned on $\hat{y}$, $\hat{x}$ is independent of π. Also, the second line shows how $\hat{y}$ really picks out either θ_1 or θ_0 and so we only need to integrate one of them. This marginal distribution is problem-specific, but as is usual, distributions are often selected so that it can be solved. Then, we would solve this integral for $\hat{y} = 1$ and $\hat{y} = 0$ and use these two values in Bayes rule above. Since we haven't defined $p(x|\theta)$ or $p(\theta)$ we have to stop here.

- Prior: The other term we need is $p(\hat{y}|\vec{y})$, which also has a marginal representation according to our model,

$$p(\hat{y}|\vec{y}) = \int_0^1 p(\hat{y}|\pi)p(\pi|\vec{y})d\pi$$

We already know from the model definition that

$$p(\hat{y}|\pi) = \pi^{\hat{y}}(1-\pi)^{1-\hat{y}}$$

and from Bayes rule that the posterior distribution

$$p(\pi|\vec{y}) = \tfrac{\Gamma(a+b+N)}{\Gamma(a+\sum_i y_i - 1)\Gamma(b+N-\sum_i y_i)-1}\pi^{a+\sum_i y_i}(1-\pi)^{b+N-\sum_i y_i}$$

We can also solve this integral, which we can write as

$$\int_0^1 p(\hat{y}|\pi)p(\pi|\vec{y})d\pi \;=\; \tfrac{\Gamma(a+b+N)}{\Gamma(a+\sum_i y_i)\Gamma(b+N-\sum_i y_i)}\int_0^1 \pi^{\hat{y}+a+\sum_i y_i - 1}(1-\pi)^{1-\hat{y}+b+N-\sum_i y_i - 1}d\pi$$

To do this, multiply and divide by the value $\frac{\Gamma(1+a+b+N)}{\Gamma(\hat{y}+a+\sum_i y_i)\Gamma(1-\hat{y}+b+N-\sum_i y_i)}$. In the numerator, put this value inside the integral. In the denominator, put it outside. Then we can notice that the integral is over a beta distribution and equals 1, so the integral disappears.

The result is

$$p(\hat{y}|\vec{y}) = \frac{\Gamma(a+b+N)\Gamma(\hat{y}+a+\sum_i y_i)\Gamma(1-\hat{y}+b+N-\sum_i y_i)}{\Gamma(a+\sum_i y_i)\Gamma(b+N-\sum_i y_i)\Gamma(1+a+b+N)}$$

One final useful property comes into play that is worth memorizing: For the gamma function, $\Gamma(x) = (x-1)\Gamma(x-1)$. Using this property on the evaluation of $p(\hat{y}|\vec{y})$ at $\hat{y} = 1$ and $\hat{y} = 0$, we see that

$$p(\hat{y} = 1|\vec{y}) = \frac{a + \sum_i y_i}{a+b+N}, \qquad p(\hat{y} = 0|\vec{y}) = \frac{b+N-\sum_i y_i}{a+b+N}$$

We now notice that we now have all we need to calculate the predictive distribution $p(\hat{y}|\hat{x}, X, \vec{y})$.

- Naive Bayes extension: For the naive Bayes extension we only need to modify the likelihood term. Using the notation from the discussion on naive Bayes above, we want to calculate

$$\begin{aligned} p(\hat{x}|\hat{y}, X, \vec{y}) &= \int_{\Omega_\theta} \prod_{j=1}^{v} p(\hat{x}(j)|\theta_{\hat{y}}(j))p(\theta_{\hat{y}}(j)|\{x_i(j) : y_i = \hat{y}\})d\theta_{\hat{y}} \\ &= \prod_{j=1}^{v} \int p(\hat{x}(j)|\theta_{\hat{y}}(j))p(\theta_{\hat{y}}(j)|\{x_i(j) : y_i = \hat{y}\})d\theta_{\hat{y}}(j) \end{aligned}$$

In other words, we have v separate marginal distributions to calculate, and the overall marginal distribution is the product of each individual one. We've made the problem much easier for ourselves, which is why someone might opt to be "naive" when doing Bayesian classification.